UNIVERSITY OF SOUTHAMPTON
Faculty of Physical Sciences and Engineering
School of Electronics and Computer Science

What Influence would a Cloud Based Semantic Laboratory Notebook have on the Digitisation and Management of Scientific Research?

by

Samantha Kanza

Thesis for the degree of Doctor of Philosophy

25th April 2018

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF PHYSICAL SCIENCES AND ENGINEERING
SCHOOL OF ELECTRONICS AND COMPUTER SCIENCE

Doctor of Philosophy

WHAT INFLUENCE WOULD A CLOUD BASED SEMANTIC LABORATORY NOTEBOOK HAVE ON THE DIGITISATION AND MANAGEMENT OF SCIENTIFIC RESEARCH?

by Samantha Kanza

Electronic laboratory notebooks (ELNs) have been studied by the chemistry research community over the last two decades as a step towards a paper-free laboratory; similar work has also taken place in other laboratory science domains. However, despite the many available ELN platforms, their uptake in both the academic and commercial worlds remains limited. This thesis describes an investigation into the current ELN landscape, and its relationship with the requirements of laboratory scientists. Market and literature research was conducted around available ELN offerings to characterise their commonly incorporated features. Previous studies of laboratory scientists examined note-taking and record-keeping behaviours in laboratory environments; to complement and extend this, a series of user studies were conducted as part of this thesis, drawing upon the techniques of user-centred design, ethnography, and collaboration with domain experts. These user studies, combined with the characterisation of existing ELN features, informed the requirements and design of a proposed ELN environment which aims to bridge the gap between scientists' current practice using paper lab notebooks, and the necessity of publishing their results electronically, at any stage of the experiment life cycle. The proposed ELN environment uses a three-layered approach: a notebook layer consisting of an existing cloud based notebook; a domain specific layer with the appropriate knowledge; and a semantic layer that tags and marks-up documents. A prototype of the semantic layer (Semanti-Cat) was created for this thesis, and evaluated with respect to the sociological techniques: Actor Network Theory and the Unified Theory of Acceptance and Use of Technology. This thesis concludes by considering the implications of this ELN environment on broader laboratory practice. The results of the user studies in this thesis have underscored laboratory scientists' attachment to paper lab notebooks; however, even though paper lab notebooks are currently unlikely to be replaced by a system of digitised experimental records, laboratory scientists are not opposed to using technology that facilitates high-level integration, management and organisation of their records. This thesis therefore identifies areas of improvement in current laboratory data management software.

Contents

List of Figures

List of Tables

Declaration of Authorship

I, Samantha Kanza declare that the thesis entitled *What Influence would a Cloud Based Semantic Laboratory Notebook have on the Digitisation and Management of Scientific Research?* and the work presented in the thesis are both my own, and have been generated by me as the result of my own original research. I confirm that:

- this work was done wholly or mainly while in candidature for a research degree at this University;
- where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
- where I have consulted the published work of others, this is always clearly attributed;
- where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
- I have acknowledged all main sources of help;
- where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
- none of this work has been published before submission
- parts of this work have been published as: Kanza et al. (2017)

Signed:...

Date:..

I hereby dedicate this thesis to Salem, I'm sure a copy of it would have made an excellent napping place for you.

"After all this time?" "Always" - JK Rowling, Harry Potter and the Deathly Hallows

Acknowledgements

I would like to thank both my supervisors Nick Gibbins and Jeremy Frey who have been absolutely amazing. Nick has shown endless patience by supervising me for two dissertations prior to my PhD, teaching me Semantic Web technologies during my undergraduate degree, and helping me apply for the Web Science iPhD in the first place; and Jeremy has been incredibly helpful, taking me under his wing when I was looking for a second supervisor and offering me multiple opportunities to get involved with other exciting projects he's worked on. Their support and advice throughout has been invaluable and I feel very lucky to have worked under such knowledgeable nice supervisors.

I would also like to thank my family for their never ending support. Firstly my father, the original Doctor Kanza, who was my inspiration for studying Computer Science in the first place, and who painstakingly proof read my entire thesis multiple times! Secondly, my mother who provided much needed emotional support and distracting phone calls when necessary. Also to Annie, who never quite understood what I was doing, but was always proud of me anyway, and who I wish had been here to see me finish; and to Mister Salem my dearly departed cat who provided cuddles and miaowed down the phone to me for support. Thank you also to Chelle Coppard for many Costa dates and work based discussions, and calming me down frequently.

A huge round of thanks also goes to all of my friends at the University of Southampton: The Web Science DTC and the Web and Internet Science Research Group who have been my university family for the last four years, and the Chemistry department for making me feel very welcome! Also to my Room4Writing group for providing motivation, support and of course, biscuits. To Alex Owen for giving me the courage to apply for the PhD in the first place, and Rikki Prince, Lisa Sugiura, Johanna Walker, Jacqui Ayling, Heather Packer, Emma Craddock, Anna Weston, Sophie Parsons, Amy Page, Mark Anderson, Rob Blair, Mark Weal, Isobel Stark and Su White for their support, advice and much needed coffee breaks! Also to Don Cruickshank for helping me with my GATE code, Susan Halford for invaluable discussions about the social theories used in my thesis, and to Simon Coles for all of his viva advice throughout the three years.

Many thanks go to Lexi Elliot supporting me throughout the entire process, and listening to me talk endlessly about my work! To my first year housemates, Steph Marshall and Rachael Evans for the late night tea sessions, and endless support even though they still pretend they don't know what my PhD is about! To Frodo Strudwick for our sibling evenings and all of his support! To Suz Coley and James Nelson for our study days, and particularly to Suz for her amazing ability to remove words from my journal paper when it was too long! To Jo Carthy and Chris Baker for amusing desk rearrangements and coffee breaks, fun nights with Sarah Jayne Jones, and their un-ending support. To Theo Chen Sverre and Geoff Howell for always providing advice, support, and hilarious distractions. To Michael Gordon (Gord) for our many Semantic Web based discussions, and Russell Frost for PhD coffee breaks! To my later housemates Sarah Dyer (for all of the midnight tea sessions) and Chloe Rose and John Capon (for the slightly earlier tea sessions) for being so awesome and supportive. To Timothy Moxon for his brotherly love and support and 5 hour phone calls, and to Jonathon Hockley, Shannon Hunt and Chantelle Mitchell for all of their support. To Peter Vaughan and Francesca Germer for the many phone calls and holiday visits. To my London buddies, Martin Warne, Hannah Sampson, Kate Robinson, James Moran and Robert Hewitt for their constant support and visits!

I would also like to thank my Chess Club friends: Jasper Tambini, Tom Wilson, Jenny Whiffin, Katy Lee, Luke Whatmough, Rhys Horlock, Alex Conway, Sonya Ackroyd, Tiana Prekodravac for listening to me talk about my PhD constantly (even though I probably bored you all to tears) and not minding when I brought my laptop on holiday with us, and Stephen Gow and Eleonore Mason for joint PhD support. And an extra special thanks to Laura and Phill Whittlesea-Clark for their never ending love and support, and Lawrence Pearman for being super supportive and keeping me sane at the end.

Another round of thanks goes to my Solent Fitness and Aerial Arts Family. I started doing Aerial Hoop and Pole Fitness there in the second year of my PhD and the studio has become my second home. I love you all, and a very special thanks to Zorena Roe, Julie Burns, Emily Draper, Hannah Bignell and Helen Jones for their support, and listening to me talk about my PhD, a lot!

A final round of thanks goes to everyone who participated in my survey, focus groups and lab observations, I would not have my data without you! I would also like to thank the EPSRC Digital Economy Programme for funding the Web Science DTC, enabling the entire PhD process!

Chapter 1

Introduction

The emergence and subsequent rise of the Web has brought with it “unprecedented communication and data exchange between scientists” (Taylor et al., 2006). Equally, technology has evolved significantly since the invention of the Web and new powerful technologies such as the Semantic Web have emerged (Berners-Lee et al., 2001). However, despite this rise of technology there is one area that remains firmly rooted behind all the rest: the paper based lab notebook, the last remaining non-electronic lab component in the scientific process (Taylor, 2006). To many, the affordances of paper still outweigh the potential consequences of not digitising their entire lab book; however with our increasingly digital era and widespread concern that important research could be lost or left behind, the idea of the Electronic Lab Notebook was born to attempt to preserve the scientific record in a more complete robust format (Bird et al., 2013).

The concept of Electronic Lab Notebooks (ELNs) have been around since the late 1990s, and numerous offerings have been made available by both industry and academia. However, thus far none of the academically produced ELNs have had a significant uptake (Hughes et al., 2004a; Sayre et al., 2017), and there is no clear top player in the ELN commercial market (Rubacha et al., 2011) despite the wide range of commercial solutions available. This thesis has undertaken an iterative investigation to understand why most scientists still don’t typically use ELNs, and whether with enough investigation and the correct user studies a solution to improve the digitisation and management of scientific research can be made. The previous offerings of ELNs have been investigated (both in industry and academia) to identify any obvious shortcomings or barriers. Three different but complementary user studies have been undertaken to characterise current lab practice and ascertain the current software usage of scientists. The conclusions of these studies combined with the results of some related research collaborations (Kanza et al., 2017) led to a conceptualisation of a three layered ELN platform, where domain knowledge and semantic web technologies were layered on top of a pre-existing cloud notebook. A prototype of the semantic layer (called Semanti-Cat) was developed and evaluated alongside further discussions around ELNs to ascertain if the extensive user studies had successfully identified what scientists wanted from ELNs, or whether there were still unrecognised barriers that would prevent them from using one.

1.1 Motivation

The main motivation behind this thesis is to improve the digitisation and management of the scientific record, both with respect to the amount that gets digitised, and the processes available with which to digitise and manage. When producing a scientific report, there is often an immense amount of different documents used to provide the information for these reports, ranging from experiment plans, observation notes, experiment data, literature notes and many more. Therefore in addition to improving the digitisation processes, it is equally important to consider how to manage this vast level of scientific research data in an efficient manner.

The scientific research process could be enhanced by being brought into the 21st Century such that it could benefit from the full power of the online digital era. In today's modern world it seems flawed that many of our scientists (and users from other domains) still rely on paper based systems. Paper has many affordances (such as its ability to facilitate quick and easy data entry) but there are so many ways that modern technology can improve a paper based system. For example, cloud based technologies enable users to access their work from multiple different locations; using Semantic Web technologies can enable documents to be marked up and embedded with metadata to facilitate more accurate ways of categorising and searching between files.

A wide variety of new systems have been created in an attempt to digitise these paper lab books; however each has placed a certain expectation on the user to adopt a new set of practices and tools which have the potential to cause disruption within the lab. Science has become an increasingly collaborative endeavour (Myers et al., 2001), and unfortunately a majority of the ELNs that are available don't facilitate collaboration; and even the ones that do, don't tie together with other collaborative tools. In addition, there has been a substantial uptake of cloud based services, such that key players in the software market including Google, Microsoft and Apple all offer cloud storage and cloud based word processing and notebook tools. This has led to a consideration of using collaborative cloud notebook tools in ELNs.

This thesis aims to investigate the requirements of the users through ethnographic studies, and to build a prototype system centred on their needs. It does not look to reinvent the wheel with another brand new system that will only appeal to a specific set of users, nor does it look to entirely replace the paper component of the research process. Instead it looks to create an ELN platform that can work alongside paper, facilitating the electronic capabilities that scientists actually want. This will be achieved by building domain based enhancements to a pre-existing widely used notebook environment and using Semantic Web technologies; to build a truly modern ELN with the required level of domain knowledge that facilitates better management and search capabilities of scientific data. It also looks to create an ELN that can present data in both a structured and interoperable fashion (Coles et al., 2013).

1.2 Research Question and Objectives

The primary research question that will be focused on is:

RQ: **Primary Research Question: What influence would a cloud based semantic laboratory notebook have on the digitisation and management of scientific research?**

The following are subsidiary research questions that elaborate on and support RQ.

RQ.1: What are the approaches and features that should be taken into account when creating an ELN?

RQ.2: What are the key processes of digitising scientific research?

RQ.3: What is an appropriate cloud environment to create an ELN with?

RQ.4: Where could an ELN fit into the current lab practice where it would actually be used?

RQ.5: How can Semantic Web technologies be utilised in this ELN environment to improve the human laboratory interaction?

Which will be worked on using the following research objectives:

RO1: To conduct a comprehensive survey of existing ELNs and ELN research (RQ.1).

RO2: To identify the current software usage of chemists (RQ.1).

RO3: To characterise current lab notebook practice through user studies (RQ.1, RQ.2, RQ.4).

RO4: To investigate the general affordances of different cloud based platforms to create an ELN environment with (RQ.3).

RO5: To experiment with the chosen platform's development environment to better ascertain its development capabilities (RQ.3).

RO6: To establish different ways of integrating 3rd Party domain specific services and semantic web services with the chosen cloud platform (RQ.3, RQ.5).

RO7: To construct a new ELN Environment to ascertain whether a pre-existing cloud based notebook with domain specific knowledge and Semantic Web technologies meets the needs of the researchers in the laboratory (RQ, RQ.5).

RO8: To conduct software evaluation focus groups to ascertain whether the semantic ELN software meets the needs of the researchers in the laboratory (RQ, RQ.1, RQ.2, RQ.4, RQ.5).

1.3 Contribution and Scope

There are four main areas of contribution for this thesis, which are as follows:

C1: ELN Market Study
C2: Initial User Studies
C3: Semanti-Cat Prototype
C4: Software Evaluation User Study

An ELN Market Research study has been conducted that collates together a vast amount of information to produce a list of over 100 ELNs, noting which ones are out there in the market, which ones are now inactive, and their costs and platform availability. The second contribution, is the results of the three initial user studies (Software usage of Chemists, Focus Groups with Scientists to establish current lab practice, Lab observations of scientists). These were all identified as gaps in current research, and necessary studies to undertake to answer the research questions laid out by this thesis. Sections of the results of these studies have been published in (Kanza et al., 2017) as part of a collaboration with BioSistemika, the Chemistry Department of the University of Southampton, and Cerys Willoughby (another PhD Student from the University of Southampton). Based on the results from these studies. a prototype piece of software called Semanti-Cat was created. This is a web application that semantically tags and marks up scientific documents. The final significant contribution are the results of the software evaluation focus groups, which facilitated refined proposals for what scientists want from an ELN piece of software. This was based on evaluating Semanti-Cat and discussing ELNs, with respect to where the PhD student participants of the study would consider using this type of software (if at all) and what tools they actually thought would have a positive impact on the digitisation of their scientific records.

Scope wise, all research objectives have been completed and some conclusions have been derived for all research questions. With respect to the proof of concept prototype (Semanti-Cat), the first iteration of the software was designed with some tweaks after a trial focus group was conducted. The evaluation focus groups led to a proposal of the next iteration of Semanti-Cat as one part of the future work section, which was considered out of scope for this thesis.

1.4 Report Structure

This report will begin by giving a history of scientific record keeping in Chapter 2, detailing the extensive ethnographic studies that have taken place in this area; followed by the technical analysis of how Computer Supported Cooperative Work (CSCW) has played a part in attempting to modernise and digitise this process. Chapter 2 will then go on to detail the ELN research that has been produced, and the offerings of industry and academia, in addition to what features and approaches are used in building them will be detailed. Chapter 3 will then discuss the overarching methodological approaches that have been used for the user studies in this thesis, alongside a discussion of the social and technical theories that have been used to interpret and understand the results of the user studies conducted in this thesis. The more refined methodology for each study is explained alongside the study write-ups in Chapters 4 and 6. Chapter 4 will then detail the initial qualitative and quantitative research that has been carried out, with an analysis and discussion of the results of these user studies. These results will then be discussed with respect to the collaborative studies that were analysed as part of this thesis. Chapter 5 will contain the technical section where Semanti-Cat, the proposed proof of concept ELN conceptualised from the three initial user studies and collaboration outcomes, will be detailed. The technical investigations of platform tools and how this prototype was created will also be described. Chapter 6 will then discuss the results of the software evaluation focus groups that took place to discuss ELNs and evaluate and iteratively design this tool. Finally, the conclusions of all research questions and proposed future work will be detailed in Chapter 7.

Chapter 2

Scientific Record Keeping

Creating and maintaining scientific records is a key part of the lab process, and one that has been in existence long before research began to focus on lab notebooks. For example, the study performed by Latour and Woolgar (2013) described some of the fundamental outputs of the lab such as graphs, charts and publications; but doesn't put a high focus on the lab notebook; showing that the concept of keeping a scientific record existed long before laboratory notebooks started receiving the high focus of academic research that they do today. It wasn't until the late 1900s that research started to focus on the concept of the paper lab notebook, Macrina (2005) takes a closer look at the importance of keeping a lab notebook, describing how entire books have been written around the topic of how to keep 'useful' or 'good' laboratory notebooks and schraefel et al. (2004) produced a comprehensive paper about the principles of scientific notebook keeping. This train of research combined with the breakthroughs in technology have progressed to research focusing on the concept of ELNs.

This chapter will begin by detailing the different aspects of scientific record keeping, and discussing the existing studies that have taken place around this area of research. Following this, it will look at the history of ELN development. The related work done in the area of Computer Supported Cooperative Work that came before ELNs were conceptualised will be detailed, discussing the transformation shift from the physical to digital, explaining why ELNs were designed in the first place, and the motivations behind this, but also why scientists continue to use paper despite the affordances of using technology instead. This will be followed by a timeline showing the evolution of ELN tools that have been created through academic research over the past 16 years, and a discussion of the adoption barriers that currently exist. The different approaches that have been taken in creating these ELNs alongside the features they commonly incorporate will then be outlined. Finally, three complementary desk based research surveys will be detailed, looking at the current state of the ELN market, ELN usage and ELN adoption barriers.

2.1 The Scientific Record

Maintaining the record is a fundamental part of scientific practice, which has been given an immense focus by scholars. It is a serially numbered collection of pages that is used to record the mental and physical activities of the scientific experiment (Kanare, 1985). It plays many important roles: It creates a legally binding record that serves as a historical artefact, protects intellectual property and provides a useful communication device to other scientists (Shankar, 2004). It also provides the ability to disseminate information and to create order (Latour and Woolgar, 2013). Existing literature will be discussed and compared to ascertain what these records perceived to be for, why they are created, and the environment they were created in. Following this, there will be a section detailing how the scientific record has been studied over the years, with a consideration of where gaps still exist in these studies to lend weight to the areas that will be examined in the user studies undertaken as part of this thesis.

2.1.1 The What

Shankar (2007) raises some important questions regarding record keeping, such as: What actually constitutes a daily scientific record? Is it solely the contents of the laboratory notebook? Or does it stretch further to contain data files, additional documents and other items associated with that lab book entry? Shankar (2007) queries the value given to these records, as it seems to be common practice to write up one's daily lab activities, consolidate these findings in conference papers and posters and then proceed to let them gather dust on the laboratory shelf never to be looked at again.

If this is the case what are these records actually for? What does the scientist who makes them day after day believe they are for? Shankar's 2004 and 2007's studies suggest that the concept of the record posed different meanings for different scientists. Two necessary characteristics were identified however, that the records be 'useful' and 'accountable' to the creator, and that they must be obtainable to the wider community at large. Shankar's 2007 study suggests that scientists create these records to produce a reliable set of information they can refer to at a later date. One of the participants of these studies described how he used 'scratch books' to scribble down his workings and equations during his experiments and then writes it up formally afterwards. He then continues by stating that he writes up everything into his lab book so that he could go back to it at a later date and it would have all the correct information in it.

A previous study conducted by Shankar in 2004 concluded that laboratory texts can be much more than a mere organisational tool. Creating these records is a very personal endeavour, and the records that were examined in this study differed greatly both in format and structure. Shankar concluded in both studies that these records only seem to be useful in the short term. Equally Reimer and Douglas (2004)'s study states that "the process of note taking is critical to scientists"; and works by Latour and Woolgar (2013) describe the role that the initial lab books and protocol books played in actually writing up a paper to be published.

2.1.2 The Why

These notes have been described as very important in the short term; for referring back to, and for aiding the scientist in question to produce publishable work. Their usefulness or lack thereof in future work past the initial publications is less clear, as they are predominantly classified as 'initially useful'. There also appears to be a call for both 'standardisation' and 'recording' within the lab environment. It could be concluded from this that scientists are both required to create these records, and feel the necessity to do so as a means of recording what they have been doing and giving themselves reliable information to refer back to, at least until the time of publication. This gives us a brief insight into *why* these records are created and their perceived usefulness, however in order to obtain a fuller understanding we need to also pose questions surrounding what their actual contents are and how they are created.

The contents of these records seem to differ from scientist to scientist, which given the aforementioned 'personal' influence on records is unsurprising. Collating together the ethnographic observations from Shankar's 2004 & 2007 studies alongside Reimer and Douglas's 2004 study, the contents of these lab book records can include the expected weights and measures for the experiment at hand, alongside annotations and observations of said experiments. In addition to this they can also incorporate tags and markers to contextualise the record, processes, references, diagrams, photos, cross references to other sections of the lab book, and masses of calculations. To further emphasise the difference in lab books, not only do they appear to differ in content, but also in organisation. According to the work of Shankar (2007) two of the participants of his study had both created different ways of standardising their data entries, that mandated a specific process of data entry and therefore a specific type of record that was created at the end of it. Temporal information was also recorded, such as the date of the experiment, and in some cases that was used to order the records and in other cases it was just recorded as a means of knowing when the experiment was conducted. This illustrates how diverse both the record and record organisational and standardisation process can be.

2.1.3 The Where

Another vitally important element to consider in scientific record keeping is the environment in which the scientists operate. Chemistry is a rich discipline with many sub domains, therefore there are a number of different types of chemistry lab setups depending on the type of chemistry that is being done. Furthermore as noted in previous studies such as the work of Hughes et al. (2004a) chemists also conduct some of their work outside the lab, for example writing up their experiments or working collaboratively with colleagues (Oleksik et al., 2014). The creation of any ELNs would need to be considered within the context of these different environments, to understand both the varying working needs and environmental conditions of these different types of chemists.

Understanding the laboratory conditions will enable a higher level of comprehension of the documents created as part of this process (Shankar, 2004). Shankar (2007) notes that creating the appropriate tools and environment for scientific record keeping is in itself part of the record creation process. Equally, when considering the wider context of how scientists could digitise their processes, it is important to be aware of the limitations of their environment. Previous studies have shown that the lab environment is a hostile one for technology in many ways, and that scientists are concerned about duplicated data entry if they have to manually take notes in the lab and then input them into a digitial system afterwards, which can result in data loss and haphazardly stored records Myers et al. (2001). However, as Hughes et al. (2004a) points out, the lab can also be a hostile space for the paper lab book, as their participants commented that they find it difficult to locate space for their lab book whilst running an experiment especially as a majority of their experiments take place in a fume cupboard which is an unsuitable environment for paper.

The concept of considering the lab environment has been at least partially adhered to in previous literature, as a majority of the studies conducted were lab observations (Reimer and Douglas, 2004; Hughes et al., 2004a; Tabard et al., 2008; Oleksik et al., 2014). These studies however generally took the form of evaluating the usage of an ELN within a lab, although some did illustrate a consideration of the environment. Guerrero et al. (2016) noted that for their lab observations of a biological ELN, they prioritised “Collaboration and Accessibility” as features to consider which ELNs they should trial in their study due to their highly collaborative environment. Oleksik et al. (2014) additionally stated that part of their analysis was to attempt to understand the work environment of the lab they were observing to comprehend how technology can support scientific activities.

2.1.4 The Who

In addition to note taking being a personal endeavour, it’s worth remembering that note taking can take different forms depending on a scientists career stage. The main participants that have been chosen to take part in the qualitative user studies (two focus groups and lab observations) of this thesis are PhD students. One of the outcomes of the software evaluation focus groups was the self differentiation that the PhD student participants placed on themselves compared to their peers in different academic career stages. PhD students arguably have a higher level of freedom to organise a large project spanning three or more years as opposed to industry projects of this length; evidenced by the contrasting approaches of the participants in the user studies of this thesis. Undergraduate students conduct much shorter pieces of work, and post docs and academic staff are typically part of a research group with varying levels of defined practices. Whereas, although PhD students belong to research groups, unless they have particularly strict supervisors, or industry sponsors, or are working collaboratively with others who have defined practices, they typically have more freedom to organise their work and use their preferred tools. This was evidenced by the results of the initial user studies, as when the participants were questioned about their use of ELNs, the PhD students who had tried using them, had been instructed to during their industry placement, whereas chose not to use them in their own PhD work.

However, despite this freedom, the concept of social hierarchy inside and outside the lab should also be considered, for this still affects note taking practices. PhD students potentially vastly differ in approaches because their supervisors follow and in some cases implement different practices, with some being more strict than others. PhD students may also take influence from other PhD students who are further along with their thesis than they are, or post doctoral researchers. These different influences across the different PhD students (particularly the chemists who worked across several different subdomains) were evident in the initial user study results written up in Chapter 4. The results demonstrated a difference in lab practice across the different disciplines and sub disciplines of the PhD students, whereby the participants who showed the strongest similarities to others in their note taking were typically part of the same research group.

Furthermore, PhD students are also preparing to produce a thesis which is a much larger document than an undergraduate dissertation, or typical journal or conference paper. Therefore the role of note taking for PhD students in science not only encompasses the expected experimental observations, data analysis and literature notes, but also involves linking back through all of their different pieces of work to eventually structure it into a full thesis. Additionally, PhD students are required to create progress reports and sometimes presentations of their work during their degree, all of which can be incorporated into the final thesis. This vast range of documents was made evident by the wide variety of documents brought along to be trialled in Semanti-Cat in the software evaluation focus groups; whereby the documents almost all differed in length, content, and format, and yet were all documents that would eventually contribute to writing a PhD thesis. Therefore the concept of improving the digitisation and management processes of the scientific record is vastly important for PhD students, as they have a wide variety of different types documents that contain information that will need to go into their thesis in one way or another, and if these records were well managed from a single system and were all available online it would improve this process.

2.1.5 The How

The area of scientific record keeping and laboratory life in general, is one that has been studied at length over the years. These ethnographic studies mostly took the form of either lab observations (Reimer and Douglas, 2004; Shankar, 2004; Hughes et al., 2004a; Shankar, 2007; Tabard et al., 2008; Oleksik et al., 2014) or interviewing scientists (schraefel et al., 2004; Oleksik et al., 2014). The lab observations were generally conducted over several months but took slightly different forms. Shankar's 2004 & 2007 studies involved spending several months in the lab recording observations. These were written up predominantly temporally to explain what had happened over time in the lab, and also included snippets of conversation and examples to put certain ideas into context. In addition to examining the why and what of the record, Shankar also investigated the how. These studies concluded that scientists still work with an unusually large amount of paper, and that seemingly some of the participants still find this to be by far the preferable method of capturing these records. The author postulates that paper is still a key part of the knowledge production system, and notes that despite concerns that introducing a digital framework to capture these notes electronically would result in replication of data capturing and integrating external data and documents; that actually seems to be the overarching aim of the scientists involved in this study.

Shankar's 2004 study was conducted to gain a better understanding of the 'nature of the record' and his 2007 study furthered that research by investigating record keeping in a research lab, and to explore the creation and use of these records. Whilst the conclusions of this study did touch upon the potential for the creation of a digital environment for which to take these notes, the original research questions had a more investigative rather than developmental aim. Reimer and Douglas (2004) took a more design based approach, and conducted their observations to inform the creation of a Web based E-Notebook. This study was conducted over a few weeks as opposed to the months taken in Shankar's studies; and whilst still of an observational nature involved discussing matters in an interview-esque fashion, but within the participants lab environment so that some observations could also be made. The observer went through the participants notes and discussed their work space and way of working with them, and this enabled a comprehensive list of the different types of notes and notebooks used within this lab to be captured. These were heavily utilised to ascertain the requirements for an E-Notebook.

These requirements, despite being the result of a lab based study, were synonymous with those of a generic Electronic Notebook (see Table 1 in (Reimer and Douglas, 2004) for full details) such as gathering information (copy/paste etc), editing information, formatting, organisational structure, ability to move content around. This suggests that not only should an ELN be able to do everything that an Electronic Notebook or generic word processing software should be able to do, but that perhaps this basic functionality is lacking in current ELNs. What is interesting is that these requirements do not mention domain specific knowledge. Reimer and Douglas (2004)'s study was conducted with a group of biologists, which also raises the interesting question of whether biologists require less domain knowledge than chemists? Or whether there is a general lack of requirement for domain knowledge?

Oleksik et al. (2014) and Tabard et al. (2008) investigated the actual usage of an ELN in the lab. Oleksik et al. (2014) observed a laboratory of biologists where all the researchers used Microsoft OneNote[1] as an ELN. This fits with the conclusions provided by (Reimer and Douglas, 2004) as OneNote is a piece of electronic note taking software that doesn't have applied domain knowledge. The screenshots of the researchers OneNote 'notebooks' illustrated that the biologists made use of the electronic features that weren't necessarily possible to achieve with paper, such as automatic linking between sections. They also used the notebooks for collaboration and it was possible for more than one biologist to work on a 'note' at the same time. However, this study also demonstrated that creating the record is still a very personal endeavour, and even using the same piece of software, different people's notebooks had been organised and maintained in different ways, and different organisations and projects required different things. The need for 'flexible' records was emphasised as well as the desire to be able to integrate different types of ELN structures.

Similarly in August 2016, Guerrero et al. evaluated an Electronic Lab Notebook 'Perkin Elmer Elements' against OneNote by trialling both in a lab of biologists for 3 months. These two notebooks were selected after a comparison of technical parameters such as ease of search and flexibility across five Electronic Lab Notebooks; both features and notebooks had been selected due to previous literature studies. The results of this study concluded that the users preferred OneNote and that it included most of the important features that had been identified by the researchers. This study shows correlations to the study of Oleksik et al., as both examined the use of OneNote by biologists; and both seemed to put a higher focus on the generic affordances of using technology over paper such as collaboration (which was another feature identified by Guerrero et al. as important in an ELN).

Tabard et al. (2008) undertook a study of a hybrid paper and electronic notebook 'Prism'. This was a system that used a special type of paper that the users could write on (like regular paper) that would then track their writing and diagrams and store a digital copy in an Electronic Lab Notebook. This was used for 9 months in an observational period and some elements were found very useful by the users such as the ability to run the software cross platform (as there were often instances where they used both Windows and Linux across one project); and the ability to tag data to organise it. There was generally positive feedback for this system, which backs up the hypothesis of Shankar that the paper component is still vital in the record keeping process. On a similar vein of assessing a prototype system, the works of Hughes et al. (2004a) also investigated

[1] `https://www.onenote.com/`

one of these, although in this instance it was a fully electronic creation. A tablet based ELN was created that users could write on like a piece of paper. This holds very similar values to the previous prototype although it uses tablet technology rather than a special type of paper. This was well received in terms of the 'neatness' and 'security' of storing the records electronically, however it was noted to be rather time consuming.

2.1.6 Limitations and Considerations of Current Literature

There has been a great deal of work surrounding ethnographic studies that either aim to investigate current practices in the lab or to either aid in development of a digital framework, or test the usage of one. However, despite this vast array of studies, there still seem to be several research gaps that have yet to be investigated; and it's also important to note that this area is ever evolving and therefore requires new analysis to keep up with the changes available in technology.

These studies tend to focus on a single scientific discipline, which doesn't necessarily highlight the specific needs of one discipline over another, and additionally doesn't capture the common requirements that are shared by different types of scientists. Which is understandable for generic all-discipline aimed ELNs; but to capture the needs of one discipline it could be worth investigating the sciences side by side to see how they differ. It is also important to note that even within the different scientific disciplines there are many different strands (e.g. in Chemistry, computational chemists work very differently to inorganic chemists (Kanza et al., 2017)); therefore even within the same discipline, different sub disciplines may have very contrasting needs and ways of working.

Furthermore, the lab environment has been observed either to see how notes are taken or to investigate the use of a specific technology, but the general blend of paper and technology used within a standard lab environment is less reported on. Additionally, commonly only one lab environment is considered in each of these studies and therefore their ELN solution is based on that one specific environment. Another area that also seems to be lacking is how these scientists use technology outside the general concept of an ELN. In order to fully understand scientist's technological needs, it is also important to understand where they absolutely require software programs for their work. Additionally, there seems to be a pattern of either conducting ethnographic studies to find out more about scientific record keeping, or identifying / building an ELN and testing that within a lab environment. Some investigations such as Guerrero et al. (2016) do conduct initial research to identify which ELNs to trial but even that process was based on evaluating ELN features based on previous literature rather than talking to the scientists themselves. These gaps will be investigated in the user studies undertaken as part of this thesis, which will be discussed further in Chapters 4 and 6.

2.2 Computer Supported Cooperative Work (CSCW)

Before ELNs had even been conceptualised, similar work around the topic of group collaboration amongst researchers had been conducted under the bracket of CSCW, which is the process of people collaborating on a piece of work with the aid of computers (Palmer and Fields, 1994). It has a wide range of focus's including developing user friendly software, and facilitating simultaneous interactions between groups, and has been playing a role in science before the concept of ELNs came to light. CSCW practitioners have also come to accept that an interdisciplinary approach between computer science and researchers from the appropriate disciplines (science in this instance) is required in order to produce a useful system for the end users.

Various CSCW systems have been created that contain elements of the ELN systems to come (Rodden, 1991). CoLab was created to facilitate group work for researchers working in the same location; allowing them to collaborate over a 'shared electronic chalkboard'. The Project NICK also contained an Electronic Blackboard, but was created to work over large-scale distributed systems. These projects were both built to enable better collaboration during meetings, but their features make up the early foundations of the lab notebook.

Given the collaborative nature of science, it is unsurprising that the work surrounding the recording and sharing of scientific data has been of great interest to the CSCW discipline. After the developments of broader, less discipline specific systems, CSCW work started to place a particular focus on the concept of ELNs, as they are seen as a vehicle to extend the functionality of the traditional paper based lab book. Additionally, centring on the 'computer supported' nature of this discipline, work has also been directed towards moving away from the physical constraints of a paper-based system towards the affordances that can be provided by digitising this process. It has been declared by Roubert and Perry (2013) that the lab notebook needs to be redefined and rethought to encapsulate the necessary collaborative nature required by research experimental work. This thesis is attempting to achieve the above and much more by taking a new approach to creating an ELN.

2.3 From Paper Lab Notebooks to Electronic Lab Notebooks

The Paper Lab Notebook (PLN) is described by schraefel et al. (2004) as a "functional record of a science experiment", something that is able to capture "text, graphics and objects" within its pages. Even today the use of paper in the lab scenario (and in many other work based situations) is still attractive. Paper records are portable, can be securely stored and do not require a power supply; as Bird et al. (2013) remind us.

There are many other advantages to using paper; indeed Cooke and schraefel (2004) recognised that the historic success of the lab book is strongly linked to the affordances of paper: the ease of data entry, its robustness, rapid access to previous material, and the ability to quickly mark pages of interest.

However, as noted by Taylor (2006), the paper based lab notebooks is the final component of the Research and Development (R&D) workflow that has yet to be digitised; and they hold just as many disadvantages as affordances. There are practical elements to consider, such as the fact that the lab is a hostile environment for paper, as described by Frey et al. (2004); and there are also many other obvious disadvantages, namely that paper cannot be searched, shared, easily backed up, or necessarily readily accessed (Frey et al., 2003). Paper based systems also do not lend themselves to collaboration; as sharing information requires scientists to make manual copies of their work and send it over to the requesting scientist. This makes it very difficult to track changes and preserve an audit trail of the work (Machina and Wild, 2013). These shortcomings heavily contributed to the conception of the Electronic Lab Notebook.

ELNs have many advantages over their paper counterparts. They provide the ability to quickly enter, retrieve, locate and share data, in addition to facilitating long term storage by being able to create backups and archives with ease (Voegele et al., 2013). ELNs eliminate the need for manual transcription and can be used by widely distributed groups (Myers et al., 2001). The development of ELNs has also been driven by the desire to audit experimental data and protect Intellectual Property (IP), as well as enabling compliance with regulatory requirements (Badiola et al., 2014).

A step further from ELNs is the concept of Semantic Lab Notebooks (SLNs) which utilise the affordances of Semantic Web technologies that will be further detailed in Section 5.4.5. These provide the ability to use open standards to expose research data as formalised metadata (Coles et al., 2013), and to link between the different sets of data collected throughout the experimental process (Talbott et al., 2005). They enable documents to be tagged and categorised to improve organisation and searching capabilities; in addition another advantage to using these technologies is to facilitate interoperability, which will allow SLNs to work with and utilise other products.

2.4 Existing ELN Tools

Over the last sixteen years there have been a number of efforts to digitise the lab book from both a biological and chemical domain perspective. Figure 2.1 is a timeline that shows the evolution of ELN tools from 2001 to the present day (Myers et al., 2001; Arnstein et al., 2002; Mackay et al., 2002; Cooke and schraefel, 2004; schraefel et al., 2004; Hughes et al., 2004b; Talbott et al., 2005; Tabard et al., 2008; Goddard et al., 2009; Borkum et al., 2010; Quinnell and Hibbert, 2010; Rudolphi and Goossen, 2012;

Walsh and Cho, 2012; Coles et al., 2013; Mohd Zaki et al., 2013; Voegele et al., 2013; Badiola et al., 2014; Clark, 2014; Harvey et al., 2014; Oleksik et al., 2014; Guerrero et al., 2016; Weibel, 2016; Google, 2017).

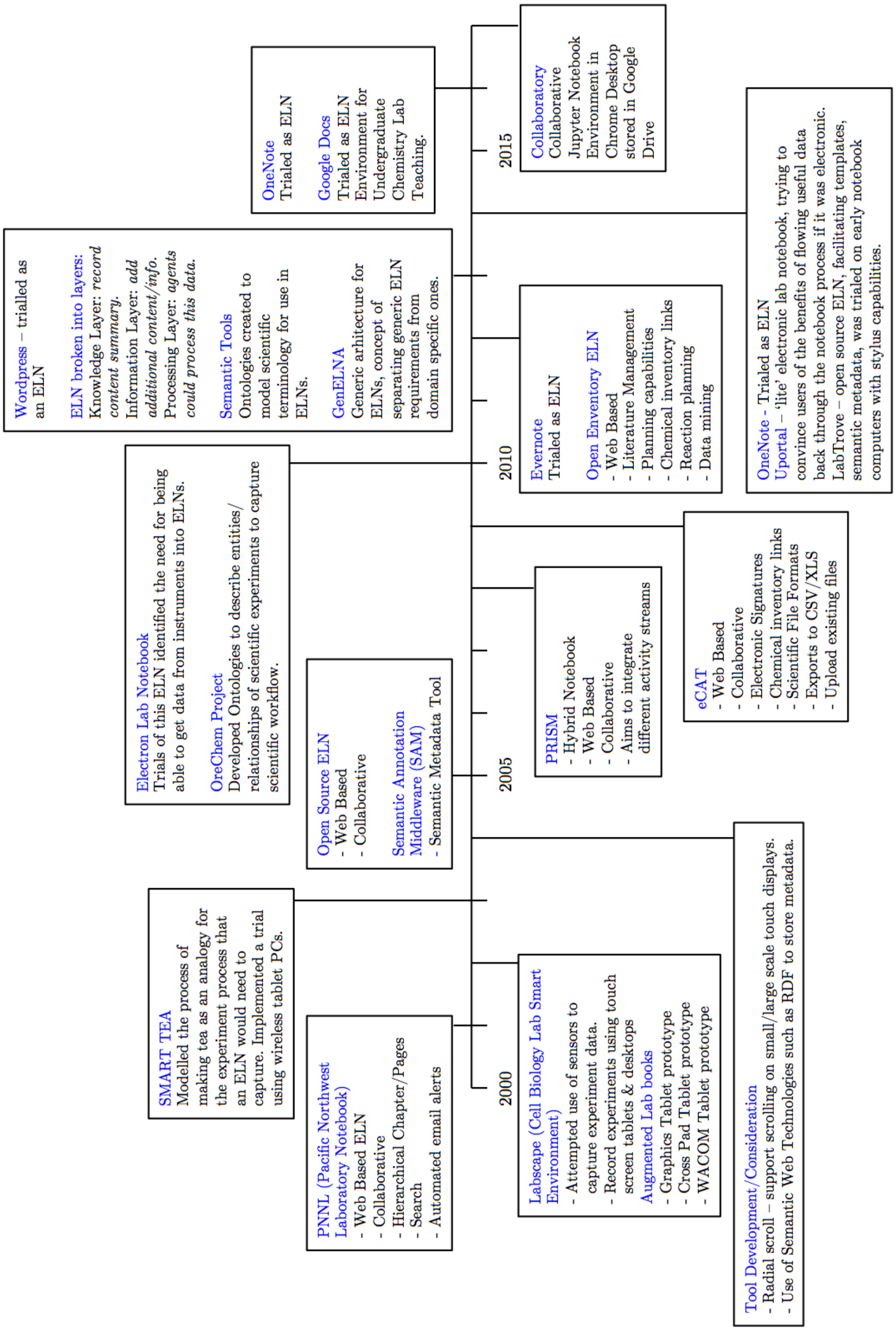

2000
2005
2010
2015
PNNL (Pacific Northwest Laboratory Notebook)
- Web Based ELN
- Collaborative
- Hierarchical Chapter/Pages
- Search
- Automated email alerts
SMART TEA
Modelled the process of making tea as an analogy for the experiment process that an ELN would need to capture. Implemented a trial using wireless tablet PCs.
Labscape (Cell Biology Lab Smart Environment)
- Attempted use of sensors to capture experiment data.
- Record experiments using touch screen tablets & desktops
Augmented Lab books
- Graphics Tablet prototype
- Cross Pad Tablet prototype
- WACOM Tablet prototype
Tool Development/Consideration
- Radial scroll – support scrolling on small/large scale touch displays.
- Use of Semantic Web Technologies such as RDF to store metadata.
Open Source ELN
- Web Based
- Collaborative
Semantic Annotation Middleware (SAM)
- Semantic Metadata Tool
Electron Lab Notebook
Trials of this ELN identified the need for being able to get data from instruments into ELNs.
OreChem Project
Developed Ontologies to describe entities/relationships of scientific experiments to capture scientific workflow.
PRISM
- Hybrid Notebook
- Web Based
- Collaborative
- Aims to integrate different activity streams
eCAT
- Web Based
- Collaborative
- Electronic Signatures
- Chemical inventory links
- Scientific File Formats
- Exports to CSV/XLS
- Upload existing files
Wordpress – trialled as an ELN
ELN broken into layers:
Knowledge Layer: *record content summary.*
Information Layer: *add additional content/info.*
Processing Layer: *agents could process this data.*
Semantic Tools
Ontologies created to model scientific terminology for use in ELNs.
GenELNA
Generic architecture for ELNs, concept of separating generic ELN requirements from domain specific ones.
Evernote
Trialed as ELN
Open Enventory ELN
- Web Based
- Literature Management
- Planning capabilities
- Chemical inventory links
- Reaction planning
- Data mining
OneNote - Trialed as ELN
Uportal – 'lite' electronic lab notebook, trying to convince users of the benefits of flowing useful data back through the notebook process if it was electronic.
LabTrove – open source ELN, facilitating templates, semantic metadata, was trialed on early notebook computers with stylus capabilities.
OneNote
Trialed as ELN
Google Docs
Trialed as ELN Environment for Undergraduate Chemistry Lab Teaching.
Collaboratory
Collaborative Jupyter Notebook Environment in Chrome Desktop stored in Google Drive

Figure 2.1: Timeline of ELN Tools

Figure 2.1 shows the shift in the approaches towards creating ELN tools. Whilst some early ELN prototypes were web based, in the early 2000s ELNs were created and trialled on portable tablet devices, such as graphics tablets or tablet PCs. This then progressed to a number of endeavours to create web based collaborative ELNs that incorporated various domain specific features. All of these ELNs were created as brand new systems with every component developed from scratch. In 2012, Walsh and Cho took a different approach by trialling existing electronic notebook software to see how it functioned as an ELN. Since 2012 Evernote, OneNote and Google Docs have all been trialled as ELNs in various settings. This illustrates a transition from trying to use a portable decide as a direct paper replacement, to providing a new online notebook environment to using existing popular notebook tools and subverting them to be used as ELNs.

This timeline also demonstrates an evolution of the ways in with researchers use semantic web technologies as part of their ELN creations. Early attempts began with trying to store semantic metadata, which later endeavours have still utilised but with enhanced methods of different layers and technologies. Additionally, new ontologies have been created to model semantic terminology and describe relationships. This shows that researchers have repeatedly seen the need for enhancing their products with semantic web technologies. This progression in approaches to ELN design and the features used will be discussed further in the following section.

2.5 Building ELNs

The three main approaches to building ELNs that will be discussed here are user centred design, ethnography and collaborative design. Different projects and studies that have utilised these techniques will be examined to see how effective they proved.

2.5.1 User Centred Design

The idea behind User Centered Design (UCD) is to enlist the help of the end users of the system, to influence how it's designed (Abras et al., 2004). A project that made great use of this technique was the Smart Tea project conducted by schraefel et al. (2004), that was part of the e-science programme. This project used the process of making tea as an analogy for a chemical experiment in an attempt to better understand the experimental recording process. Domain experts worked closely with the design team to model the process required, and the conclusions drawn from these methods were that they were valuable in creating a system that will actually be used.

Another project that utilised the user centred design approach was the eCAT ELN[2] (Goddard et al., 2009) which was developed by Axiope in close collaboration with lab

[2] http://www.researchspace.com/electronic-lab-notebook/index.html

scientists. This also received a positive response from users in terms of building a useable system. The comments made as part of the eCAT case studies[3] highlighted the features that the users liked and had found the most useful. These identified features that had been specifically built based on the UCD and the users criticisms of the previous systems.

It is vitally important to consider the users' needs when building any system. This is an opinion shared by Kihlén (2005) who conducted a user satisfaction survey of a R&D ELN used by employees of Bitvitrum (a pharmaceutical corporation spinout company). This survey concluded that the "key ingredient for successful implementation of ELNs is a user-friendly software"; which UCD can help provide. This illustrates that UCD had a positive impact in identifying what features would actually be valued by the user.

UCD techniques were also utilised by Reimer and Douglas (2004) to create their web based E-Notebook. They devised a set of research scenarios as part of a usability study to analyse how well a set of biologists could complete tasks with the software they had access to. They used their results to aid in the creation of their Notebook, and when their software was tested against the existing systems, the results showed a favourable comparison. These examples therefore show that in both proof of concept projects and an actively running ELN that UCD can be a very useful tool in understanding the users processes and how best to shape new systems to fit them.

2.5.2 Ethnographic Studies

Hand in hand with user centred design is the ethnographic approach. Ethnography involves the researcher taking either an active or passive participation in people's daily activities centred around the topic of research. It typically includes interviews (either formal or informal) collecting information and observing people's actions in an everyday context (Hammersley and Atkinson, 2007). Section 2.1 detailed the findings of previous ethnographic studies and the motivations and methods behind them, illustrating how they have been used to inform researchers about the scientific record keeping process. This section will look at the benefits of conducting these ethnographic studies.

A study conducted by Oleksik et al. (2014) involved lab observations of eight individuals to study their movement and work practices. They also interviewed said participants extensively to discuss how they ran their experiments in the lab and what collaborative communication methods they used. This led to identifying crucial needs of the user including "flexibility and fluidity of editing records" and their requirements for a persistent consistent system.

The Smart Tea project (schraefel et al., 2004) also led to interviewing scientists to identify the components of the experimental process (similarly to the original method

[3] `http://www.researchspace.com/electronic-lab-notebook/blog/site/?cat=16`

of cataloguing the process of making tea) to ascertain the system requirements. Intensive interviewing is deemed a useful technique by Consolvo et al. (2002), as it allows the interviewer to learn about the user's work in their own words. It also allows them to establish a rapport with the users, which can help encourage the user to be more forthcoming, and makes it easier to use additional ethnographic techniques with the user's cooperation. This demonstrates that an ethnographic approach can be useful in determining the needs of the user and the processes that need to be considered.

As a further part of the e-science programme Hughes et al. (2004a) also did a great deal of additional research around this topic. They noted that ethnography revealed the wider set of interactions that existed in an experimental environment, which helped them to appreciate the context in which the lab book is viewed. Not only does it enable the researchers to better understand the complex communications and interactions that exist in a lab environment; it allows for observation of users every day actions to capture common behaviours with respect to the lab as well as their obvious requirements. This is especially important for this thesis, as it is looking to capture the process of the lab, not just to merely provide a method to output their experiment results.

2.5.3 Collaborative Design with Scientists

The Smart Tea project (schraefel et al., 2004) also focused on collaborating with chemists to produce a digital recording system for the lab that permitted chemists to interact with it in a way akin to how they interact with their lab book. The chemists were involved in each developmental stage and proposed changes they deemed necessary at the change reviews. This allowed the developers to gain instrumental insight into the needs of the user, and enabled a mesh of developers and scientists to produce a functional tool that captured the necessary features.

The oreChem project (Borkum et al., 2010) which was funded by Microsoft to investigate designing and developing semantic chemistry infrastructure also makes use of collaborative development techniques. It involved collaboration between computer scientists and chemists to develop and deploy ontologies for describing the entities of a scientific experiment.

It is evident that the people best placed to help design an ELN are the scientists; who are not only potential users, but also who understand how the experimental process works. This has been illustrated by the examples given, which demonstrate how the domain experts (chemists) can advise the developers (computer scientists) on how to create better systems. It also demonstrates how by combining these different skills sets; it is possible to bring the best elements of both disciplines in a collaborative sense to produce a more comprehensive system.

2.6 Features of ELNs

Having looked at the different approaches to building ELNs, this section now discusses the different features that have been and could be incorporated into an ELN.

2.6.1 Collaborative

Science is an increasingly collaborative endeavour, and therefore it is unsurprising that attempts have been made to design collaborative ELNs; as demonstrated in Figure 2.1. The Pacific Northwest National Laboratory worked with Myers et al. (2001) to produce a collaborative based Web Notebook primarily aimed at biologists as a proof of concept project to investigate its popularity. This project was unfortunately shut down in Winter 2007 due to funding running out[4]. However, it did produce a positive result; it was successfully used as a collaborative tool in a set of education and research based settings.

A study that was part of the CSCW conference in 2014 investigated how ELNs affected collaborative work between scientists. (Oleksik et al., 2014) investigated the use of OneNote[5] (a Microsoft collaborative notebook) as an ELN. It was used as a collaborative ELN: scientists shared their personal notebook with each other, and shared notes were created to support discussions and note taking in meetings. The files were stored on a centralised server that all members of staff had access to. The results from this study found that participants used ELNs for synchronous and asynchronous collaboration with group members. Shared notebooks were used to collaborate successfully during group meetings; and that researchers would send links to their notebooks to their research leader to gain rapid feedback which enabled a much quicker more effective response.

Given the substantially more collaborative nature of Web 2.0 and the increasingly collaborative nature of science, it seems only natural that these features should have worked their way into ELNs. However, even a few years ago it was noted by in Quinnell and Hibbert (2010) that there weren't a great deal of accessible ELNs that facilitated collaboration; which explains why a generic Electronic Notebook (EN) was used to test out ELNs as a collaborative device. However, the two studies depicted here illustrate that collaboration can be used effectively in ELNs and that it facilitates faster and easier sharing and the ability to communicate faster, and simultaneously irrespective of location. This combined with the fact that collaborative tools have been increasing in their usage (a study from 2013 put Google Apps users at 50 million[6]); they have become an important focus of digital tools, and a highly useful characteristic of ELNs.

[4] `http://collaboratory.emsl.pnl.gov/software/software-notice.shtml`

[5] `https://www.onenote.com/`

[6] `http://blog.bettercloud.com/google-apps-stats/`

2.6.2 Platform Independent and Cloud Based

As described further in Section 2.7 there are also a wide number of commercial ELNs available. However, as noted by Rudolphi and Goossen (2012), despite how these solutions should theoretically be attractive to research groups due to the additional functionality they provide, there are a lot of barriers to this. Most commercial ELNs (as supported by the investigation described in Section 2.7.2.2) are costly due to licensing restrictions and often rely on enterprise level database infrastructure. Given the closed structure of proprietary solutions, users are beholden to the software suppliers to maintain and update the software. Therefore, the notion of a platform independent ELN could avoid some of these issues.

There are two main ways that an ELN could be platform independent. It could either be a downloadable application that was made available for different operating systems, or it could be Web based. The virtues of web based tools are discussed by Rudolphi and Goossen (2012) who note that a Web based ELN can be available via any Web browser with an internet connection. Cloud based applications have been described as web based applications with notable enhancements (TechRepublic, 2012). The main enhancement that has been detailed is the notion of being able to store all of your data in the cloud. Therefore, a cloud based ELN would permit the user to access their ELN and experiment data from any computer with Internet access.

In 2014, Clark presented the notion of creating a user interface for mobile devices that provided a large array of chemistry based products with support from a number of cloud based services to enhance functionality. The aim was to produce a full ELN environment for capturing chemical reactions. This was described as a unique product based on the high-level chemistry domain knowledge incorporated in it. It was also noted that a 'new generation' of computing platforms are able to play a role in the environment of cheminformatics. This illustrates how researchers are starting to utilise the benefits of cloud services, in addition to making use of the wide variety of different mobile devices that are available, so that ELN environments can be accessed anytime, anywhere, from any type of device with internet access.

2.6.3 Rich Semantics

Another area that has commenced exploration in the early 2000s was the notion of an ELN that can "act as a component in a wider encompassing semantic system" (Talbott et al., 2005). These researchers used the ELN client developed by the Pacific Northwest National Laboratory to investigate the use of Semantic Web technologies within an ELN (Myers et al., 2001). This ELN focused on storing its data in underlying repositories, exposing its metadata in RDF and integrating with other systems to produce and consume metadata. Whilst this ELN might have been discontinued, the work done concluded that taking the data from ELNs and exposing it as standard metadata (which other ELNs such as Labtrove have also been focusing on (Badiola et al., 2014)) is an important step in bringing the ELNs into the wider world of the "Semantic Web and knowledge grids". In addition, they felt that creating ELNs that are capable of producing a shared record with underlying semantics will be a "key enabler of next generation research". This supports the one of the motivations behind this thesis, which is to bring the full power of the 21st century technological capabilities to lab research.

The CombeChem project from the University of Southampton has also investigated the use of Semantic Web technologies within ELNs. An ontology was produced by Hughes et al. (2004b) that encompassed the main phases of an experiment to record the links between the data gathered during the experiment, what that data means and the relationships between the two. This gave way to a great potential for additional functionality such as making inferences about the types of experiments that were logged, and creating valuable links between experiment outcomes and their final reports. Additionally, the oreChem project (Borkum et al., 2010) also designed an ontology to describe the entities of experiments in order to provide machine readable data regarding methodologies and results. Clearly it was felt that the richer data formats provided by the Semantic Web would be able to facilitate a more detailed description of experiment data. Additionally, the more machine readable the ELN data is, the more it becomes interoperable which is one of the huge benefits of incorporating Semantic Web technologies into ELNs.

There has also been some work done in the area of Semantic Notebooks as a general concept. Drăgan et al. (2011) focuses on some of the affordances of the Semantic Web; namely how to interlink important information and how to design interfaces to support the existing workflow of the user. This work facilitated a study of users to see whether they preferred SemNotes (a new Semantic notebook) or Evernote, which produced favourable results with respect to SemNotes. This demonstrates that these additional features can even outperform a product as popular as Evernote. This survey was conducted before Evernote started incorporating Semantic Web technologies within their context booster[7], that provided links to additional content on the web to enrich users notes. This shows that both domain specific and general ELNs have seen the value of Semantic Web technologies, as have both commercial and academic all purpose ENs.

However, despite these attempts, as evidenced by the results in Section 2.7 there is no Semantic chemistry notebook available in the market. Additionally, despite Figure 2.1 illustrating the work done with respect to Semantic Web ELNs, most of this work has been disjointed from actual electronic lab notebook software, such as creating ontologies or a semantic middleware platform that could be used alongside an ELN. This means that despite the affordances of Semantic Web technologies, and despite researchers consistently recognising the part they could play in an ELN, they have yet to become part

[7] https://discussion.evernote.com/topic/39748-context-booster-your-knowledge-assistant-for-evernote/

of the recognised package of an ELN. This is an area that has been investigated in great detail on an academic level (Hughes et al., 2004a,b; Coles et al., 2013; Borkum et al., 2010) and clearly has some worthwhile affordances as described in this section. Therefore they are definitely worth including, but how to include them in a more successful way will need to be investigated further.

2.6.4 Domain Specific

A domain specific ELN whereby there is one ELN per scientific discipline isn't necessarily the best idea, as noted by Kihlén (2005). He warns against using an ELN as a tool to fix discipline specific problems, rather than as a tool to bridge between projects. Researchers such as Voegele et al. (2013) clearly agree with this line of thought, as they have worked to produce an Open Source universal ELN using WordPress.

There were some attempts to use a EN (Evernote) as an ELN, however Walsh and Cho (2012) concluded that the users felt that despite the affordances of using Evernote over a paper notebook, it was lacking in the domain knowledge. It was described as "simple and practical for some laboratories, but for others it does not offer features specialised for fields such as biology chemistry or quality assurance/quality control". This is unsurprising as different scientific disciplines may have contrasting needs for diagrams, equations, or scientific notation capabilities. This suggests that there could be some merit in using an existing system but that it would require additional domain specific features.

With the exception of trialling a few electronic notebooks such as Evernote and OneNote, a majority of the work done with these ELNs have involved using existing systems, they have all been newly created systems. A majority of them have been explicitly looking to fill a specific gap (for example the work done by Kihlén (2005)) or to reinvent the wheel by creating a new general all-purpose ELN. There are disadvantages to having ELNs as a standalone system as noted by Talbott et al. (2005), and thus far, none of these solutions seem to have had a great uptake, and several have been discontinued due to lack of funding such as (Goddard et al., 2009)'s eCAT.

This suggests that currently the right balance has yet to be struck between domain knowledge and the ability to use an ELN for more than one purpose or discipline. It is clear from research discussed above, that domain knowledge is important and must be included in any ELN; as even a popular general-purpose notebook like Evernote was found lacking in this area. However, whilst the notion of an all-purpose ELN is popular based on various ones being developed, they evidently haven't been created in a way that took off either; and trying to make them all purpose potentially lost some of the specifics that were required. Therefore, an ELN created off an existing notebook system but with additional domain specific knowledge could prove popular.

2.6.5 Build upon an existing Electronic Notebook

A majority of the previous and existing ELNs (as illustrated in Figure 2.1 involved creating a brand new (and often complex) architecture (Mohd Zaki et al., 2013)'s. Even those that built upon existing architectures still created a new notebook system to overlay it. Prime examples of this are Talbott et al. (2005)'s ELN (a collaborative web based ELN) which is based on the Scientific Annotation Middleware (SAM) architecture developed by Myers et al. (2001). However, all of these involved creating at least the notebook side from scratch, whereas the approach suggested here is to build upon a pre-existing notebook environment. There are a number of reasons for this: Firstly, to utilise the collaborative cloud based features that were clearly popular in previous ELN work. Secondly, having investigated the features incorporated by ELNs, electronic notebooks and indeed semantic lab notebooks, there is a large degree of overlap; thus meaning that it should be possible to build an ELN environment upon a Notebook.

This is backed up by the work done in recent years by (Reimer and Douglas, 2004) who concluded after observing a group of biologists that the main features required by a web based electronic notebook were the features you'd expect to see from a generic electronic notebook, and the shift to trialling existing electronic note booking software as ELNs undertaken by (Walsh and Cho, 2012; Oleksik et al., 2014; Weibel, 2016; Guerrero et al., 2016). The diagram below details a model of features that have been detailed in ELN research work broken down into the three layers, (Frey et al., 2003, 2004; Hughes et al., 2004a; schraefel et al., 2004; Bird et al., 2013; Talbott et al., 2005; Voegele et al., 2013; Reimer and Douglas, 2004). This model forms the original basis for the model detailed in Figure 5.1 in Chapter 5.

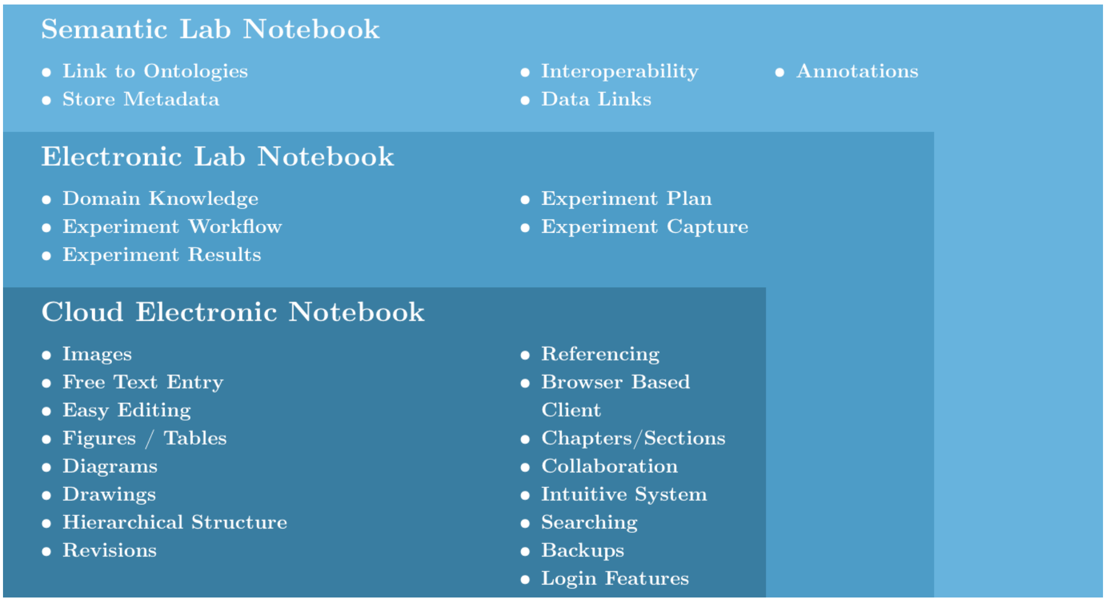


Figure 2.2: Intial Model of the Three Layered ELN

The advantage of building an ELN on an existing system is that it would allow us to make use of the overlapping features demonstrated in Figure 2.2. In addition, any ELN created upon an existing system would already have an established user base and third parties would support the main notebook elements. If there was a way to make the ELN extensions open source or allow developers to collaborate on them, then this would provide even more advantages. Then, even if the original developers ceased support and development, new parties could take them up. It would also provide a greater focus on the specific domain based features.

The affordances of the features discussed in this section will be taken into account when creating this ELN. It will be created upon an existing cloud based notebook platform with collaborative capabilities, and will incorporate both domain specific knowledge and semantic web technologies. The techniques of user centred design, ethnography and collaboration with scientists will also be used to assist the development process. How this will be achieved will be detailed in sections 4, 5 and 7.

2.7 Desk Based Research

This section details three complementary surveys that look at the current state of the ELN market, ELN usage and ELN adoption barriers. As part of this thesis, a study was undertaken in line with Research Objective 1 (see Section 1.2) to investigate what ELN tools exist in the market. Section 2.4 looked at the ELN tools that had arisen from academic research, but this study investigated what tools had actually been created and released to the public. This study was conducted to gain a better understanding of what ELN offerings were commercially available to gain a better understanding of any potential gaps in the market and to provide more insight when investigating both ELN usage and their adoption barriers. The other studies detailed in this section are ones that were undertaken by collaborators in the work done by (Kanza et al., 2017), and analysed as part of the work completed for this thesis. These comprise of a survey conducted by BioSistemika to look at the current usage of ELNs and the perceived barriers to their adoption, and a complementary study conducted by the Dial-a-Molecule group from the University of Southampton who also conducted a survey investigating the barriers to adopting an ELN as part of their iLabber Pilot Project. These two studies are detailed first, followed by the ELN market study; the three studies combined offer collective insights into how ELNs are currently perceived.

2.7.1 ELN Usage and Adoption Barrier Surveys

Despite the many ELN tools that have been created, there is still a limited uptake in academia of ELNs. BioSistemika's 2015-2016 survey of ELN usage identified that only

7% of their participants were actually using an ELN (detailed in Figure 2.3); only 10% of participants said they were actively looking for an ELN and 62% answered that they would think about it in the future. A 2017 study conducted in America, also identified a very low level of ELN usage at top American universities (Sayre et al., 2017), suggesting that ELNs are still not a popular technology.

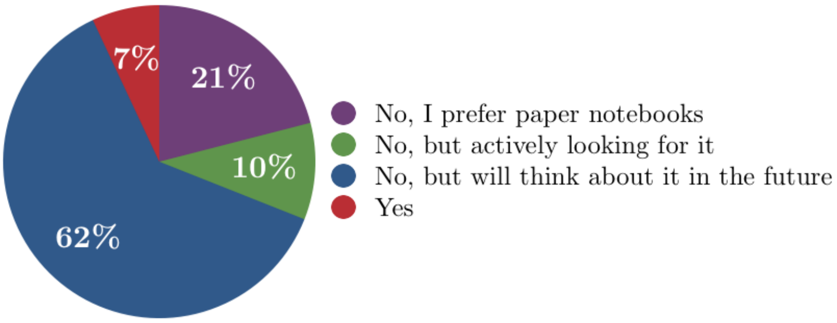


FIGURE 2.3: BioSistemika's Survey of ELN Usage from October 2015 and February 2016, adapted from (Kanza et al., 2017)

Therefore, despite the affordances of storing ones work digitally, there are still many adoption barriers to using ELNs. A 2011 survey conducted by the Dial-a-Molecule iLabber Pilot Project at the University of Southampton, asked 169 participants about potential issues that would prevent them from using an ELN and their responses are summarised in Table 2.1 under the main categories of: cost, ELN attitude, ease of use, ELN access, Data compatibility and other (Kanza et al., 2017). These barriers will be explored in more detail later on in this chapter to explain why participants identified these issues. The main barrier categories will also be referenced in Table 4.3 to show which user desired features identified from the user studies in Chapter 4 could mitigate these barriers.

2.7.1.1 Cost

This is a significant barrier to ELN adoption (Bird et al., 2013; Goddard et al., 2009). Biosistemika's 2015 ELN Study detailed in (Kanza et al., 2017) identified 'Limited Budget' as their highest adoption barrier, and as Table 2.1 illustrates, a significant number of the participants in the Dial-a-Molecule study identified cost as a barrier. Cost doesn't just represent an initial financial outlay to pay for a system, there are also often hidden costs, such as time costs for users to learn how to use a new system and troubleshoot any issues. Figure 2.5 detailed later in Section 2.7.2.2 shows that in the market today only 23.9% are open source or have free versions of their software, suggesting that a majority of the current ELN offerings have the deterrent of cost. This conclusion is supported by the results of BioSistemika's survey when they asked their respondents how much they would be willing to pay per user per month for an ELN (Kanza et al., 2017). These results showed that none of the participants were willing to pay any more than $50 per user per month, which was half the number of participants who said they'd only be willing to use free software. These findings suggest that cost plays an important part in ELN adoption, and this would need to be seriously considered in the development of a new ELN tool.

2.7.1.2 ELN Attitude

20% of participants in the Dial-a-Molecule survey said that they thought adopting an ELN would only make sense if the entire department adopted it, with 11% sharing the view that postdocs and students would resist ELN adoption. This relates closely to the Unified Theory of Acceptance and Use of Technology (UTAUT) theory that social influence and facilitating conditions play a part in ELN adoption (Venkatesh et al., 2003). This theory hypothesises that the degree to which one's superior believes that an individual should be using a technology, and how much the individual feels that this technology will be well supported has an effect on whether they will adopt said technology. One of the ways that an entire department would be likely to adopt an ELN would be if a senior member of that department was behind the decision. Additionally, how well an ELN would be supported both technologically and within the departmental organisational structure would effect the extent to which both department heads and students / postdocs would object to adopting an ELN.

Barrier Category	Barrier Explanation	% of 169
Cost	Up-front costs and licensing fees	74
	Additional infrastructure costs (e.g computers)	27
	Future development and costs of application	90
	Ongoing costs of the system	93
ELN attitude	Only makes sense if the whole department adopts it	20
	Belief that students/post docs would resist adoption	11
Ease of Use	ELN was too difficult to use	22
	Does not capture the right information for me	7
	Hard to capture some types of information in ELNs	80
ELN Access	Duplicated data entry across the lab / write-up area	74
	No easy access to appropriate hardware in the lab	12.5
Data Compatibility	Data will be tied into a commercial package	84
Other	Other	11

Table 2.1: Categorised Barriers of ELN Adoption from the Dial a Molecule iLabber Pilot Project: Potential Uses of ELNs in Academia Survey from September 2011, adapted from (Kanza et al., 2017)

2.7.1.3 Ease of Use

UTAUT hypothesises that the effort expectancy of a technology (how easy it is to use) impacts on its adoption (Venkatesh et al., 2003), and in keeping with this, ease of use has been identified in the Dial-a-Molecule study as an adoption barrier. 80% of participants thought that it would be difficult to capture certain types of information in an ELN. Even in today's technological age, paper still has many ease of use affordances over technology; it facilitates completely free entry of text and data in whatever format the user desires (e.g scribbled notes, diagrams, flexible use of the space on the piece of paper). Early Biology ELN prototypes used graphics tablets to simulate that free pen to paper text entry (Mackay et al., 2002), and whilst these were acknowledged as expensive solutions, the affordances of this free text entry was popular with the users who trialled these systems. Tablet based solutions are definitely still a costly endeavour today, but these earlier and recent findings do suggest that not only is ease of use vitally important to users, but that perhaps a full replacement of paper would remove some of this ease of use or result in inflated costs. Thus some middle ground may need to be found when conceptualising a new ELN solution.

2.7.1.4 ELN Access

74% of the respondents to this survey were concerned about duplicated data entry, with 12.5% concerned that there would be no easy access to appropriate hardware in the lab (tying into the point above that having the right hardware in the lab isn't always feasible). Lack of suitable hardware in the lab suggests that scientists would then need to input their data into an ELN after already creating the entries in their Paper Lab Notebook. However, comments from a follow up survey also conducted by the Dial-a-Molecule group (Kanza et al., 2017) revealed anxieties towards taking one's laptop into the lab due to being concerned about damaging or contaminating the laptop, and the lack of space in the lab to rest the laptop. This suggests that lack of hardware is a complex issue as seemingly even if more hardware were available there would still be concerns about taking it into the lab in the first place. Thus following that, it would be difficult to facilitate a system where necessary ELN interaction could take place inside the lab.

In a survey conducted by the Public Research Consortium in 2010 it was revealed that even though researchers understand how important it is to be able to access their data, they do not find it easy to access. De Waard (2016) suggests that this is due to fragmented data storage opportunities, and the fact that often there are not clear data management practices to follow. These factors all lead back to the performance expectancy element of UTAUT (Kanza et al., 2017) as arguably scientists would be less likely to perceive an ELN as something that would enhance their performance it if was doubling their work load, or if they didn't feel that it would improve their performance in the first place. Similarly, in order to address the ease of use adoption barriers, a middle ground between Paper Lab Notebooks and Electronic Lab Notebooks may need to be found.

2.7.1.5 Data Compatibility

The final categorised adoption barrier highlighted by this study is data compatibility. As Figure 2.5 demonstrates, 88.6% of the current ELNs in the market are commercial systems (e.g. fully commercialised or have a free and paid for version), and therefore are likely to have their own data formats. Comments from the Dial-a-Molecule survey highlighted concerns that participants were worried about being tied into commercial data formats or ending up in a situation where they were unable to get to the raw data they had input into some software. To mitigate this, a new ELN platform could make use of common data formats both in terms of how the data is stored and the export capabilities.

2.7.2 ELN Market Survey

Following the extensive research performed surrounding ELNs, this section will summarise the existing commercial and open source ELNs. The methodology of how these were investigated will be detailed followed by the findings which will be presented graphically. This study is an updated version of the one published in (Kanza et al., 2017).

2.7.2.1 Methodology

A simple searching strategy was initially adopted to investigate which ELNs exist in the market. The first step was to use Google to ascertain if there were any useful sources that could be used. Four main sources were identified from this search, two wiki based websites (Atrium Research, 2018; LIMSwiki, 2018) and two research papers (Taylor, 2006; Rubacha et al., 2011). There was a degree of overlap between the sources, so a unique list was created that contained each ELN from these sources, in addition to a few others that were found amid searching. From these sources, 122 Electronic Lab Notebook products were identified and each ELN on this list was then further investigated. The full list of ELNs, with links to a full spreadsheet of further information can be found in Appendix A.

Whilst these sources contained a wide variety of ELNs, the two journal papers weren't very recent and (LIMSwiki, 2018) also listed several inactive ELN vendors. Subsequently, the next investigation was to split out which ELNs were no longer active, by searching to see if it was possible to find an active website detailing this product under the same vendor as it was originally listed under, and if not a separate search was performed to note what had occurred. Additionally, ELNs that didn't meet with the standards defined by the Collaborative Electronic Notebook Systems Association (CENSA) standards for ELNs (ConsortiumInfo.org, 2013) were also eliminated; which led to 34 being removed from this list. This searching strategy proved to be thorough, although that was potentially due to finding sources that listed multiple ELNs.

The remaining 88 active ELNs were then categorised by the disciplines/domains they supported. Given the chemistry focus of this paper, this was a given category. Several ELNs also described themselves as being created either for the biological, pharmaceutical or life sciences. Research and Development (R&D) and Quality Assurance / Quality Control (QA/QC) were categories that were listed by both (Atrium Research, 2018) and (Taylor, 2006) and seemed to be sensible additions. There were also multiple ELNs that defined themselves as supporting more than one discipline (e.g. chemistry and biology). These were then defined as Multidiscipline. The final category is Generic for ELNs that describe themselves as general ELNs to meet all lab needs. The ELNs were then put into these categories based on the following factors:

- **Chemistry** - either specifically designed for chemistry, or had a chemistry specific version of the ELN.
- **Biology** - either specifically designed for biology, or had a biology specific version of the ELN.
- **Pharmaceutical** - either specifically designed for pharmaceutical sciences, or had a pharmaceutical specific version of the ELN.
- **Life Sciences** - either specifically designed for life sciences, or had a life sciences specific version of the ELN.
- **Research & Development** provides R&D tools or are tailor made for R&D.
- **Quality Assurance / Quality Control** - provides QA/QC tools or solutions specifically designed for QA/QC needs.
- **Multidiscipline** - ELNs that cater to more than one of the categories detailed above. Note: ELNs that have separate versions for specific disciplines are not included in here; they are in the categories for that discipline alone.
- **Generic** - this is similar to the category above, but is for general ELNs that aim to support all lab needs.

The ones that supported chemistry were then broken down into which platforms they supported, Windows, Mac or Platform Independent. They were also examined to ascertain whether they were commercially licensed, Open Source or free. The outcomes of these categorisations will be discussed in Results Section below.

2.7.2.2 Results

This section details the results of the survey of ELNs. As illustrated by Figure 2.5 of these 33 chemistry ELNs, 45.5% are web based, a further 12.2% are platform independent, with the rest either being Windows or Mac only, or not specified. This fits with the progression shown by Figure 2.1 that ELN tools are moving to be web based, or at least available on all platforms. Over 50% combined being web based or platform independent is consistent across the entire ELN domain, further reinforcing this.

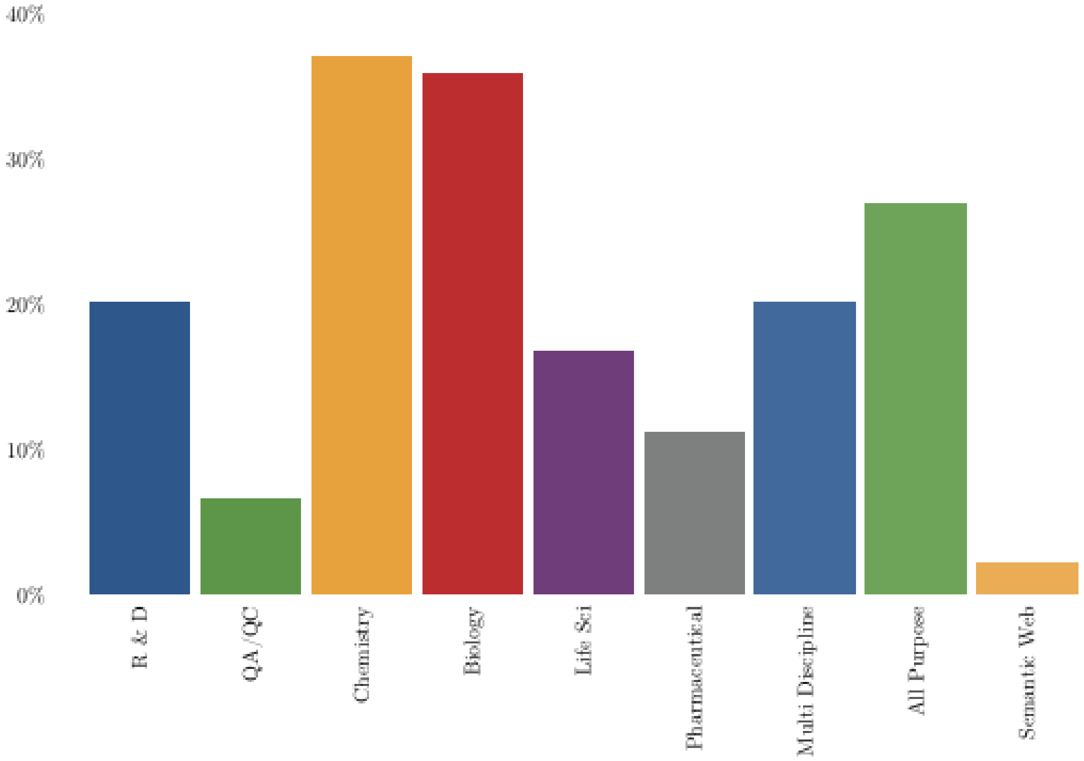


Figure 2.4: ELN Market Study - Active ELN Vendors, adapted and updated from (Kanza et al., 2017)

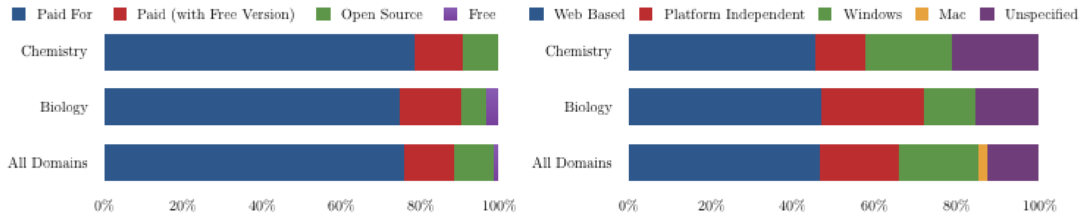


Figure 2.5: ELN Market Study - Licensing and Platform Information of Active ELN Vendors, adapted and updated from (Kanza et al., 2017)

With regards to licensing, of these same 33 chemistry ELNs, only 12.1% of these offerings have free versions (which are typically versions with reduced functionality or adverts), although some offer free versions for academics. Additionally, there are 9.1% of open source ELNs. These are similar figures to the overall ELN domain, although there is one free offering for biology ELNs. This feeds into the adoption barrier of cost, as a majority of these solutions are expensive commercially licensed products. Furthermore, open source offerings can have a concern attached to them that there won't necessarily always be technical support for these products, relating to the facilitating conditions variable of UTAUT.

Looking at both the platforms and the licenses combined, for the 33 chemistry ELNs, there are 3 ELNs that are web based and have free versions, and a further two that are Open Source. Therefore out of an original 122 ELNs only two fall into the category of: A chemistry ELN that is both unconditionally free (without the dependency on additional costly software, adverts, or the need to be in academia) and web based.

As for the notion of a SLN available in the market, only two of these make use of Semantic Web technologies. MyLabBook[8]) makes use of Semantic Web technologies by linking to an experiment ontology, and this is a biology specific ELN, and CEFF [9] uses semantic metadata tools. Some generic Electronic Notebook products have tried to make use of Semantic Web technologies, (such as Evernote with their context booster)[10] although that was unfortunately shut down due to lack of funds.

2.8 ELN Survey Discussion & Summary

The results detailed in Figure 2.3 illustrate that there are people looking to use ELNs or actively considering it, but despite this there are still barriers to adoption. A lot of the ELNs out there in the market are commercially licensed and require complicated setups, and even the rare free ones mostly have some form of drawback such as adverts or a dependency on other software. This proprietary nature is a barrier to their adoption as there is a high overhead to switching to these systems, and cost has been identified as a significant factor in adoption. There are still about 40% that aren't platform independent and there are concerns about data compatibility and portability issues. There are also concerns with respect to whether it would be feasible to use hardware in the lab to fully facilitate recording the scientific record digitally, but that without this there may be an overhead of duplicated data entry, and either way there are also feelings that it wouldn't be possible to capture all of the information types required in an ELN.

Some of the initial conclusions in this section have identified that perhaps a middle ground between removing paper entirely, and losing some of the affordances and ease of use of paper by digitising everything needs to be conceptualised. Additionally, it has become apparent that whilst ELNs aren't overly popular, they have a number of generic

[8] http://www.mylabbook.org/

[9] http://cerf-notebook.com/

[10] http://www.contextbooster.com/

features in common with the much more widely used electronic notebook solutions. Therefore building an ELN on top of an existing notebook system, ideally a cloud based one so that it could easily be used on any mobile or desktop machine (Tabard et al., 2008), that used existing formats and was already widely used could be a potential step forward. Similar endeavours have already been attempted by the works of (Walsh and Cho, 2012; Oleksik et al., 2014; Weibel, 2016; Guerrero et al., 2016), however these studies looked at using an existing electronic notebook tool as it is to see how it functioned as an ELN. Layering domain knowledge and semantic web technologies on top of this existing platform could facilitate a familiar easy to use platform, with the required domain knowledge that is currently lacking from electronic note booking software, and the ability to add metadata and additional knowledge to the records using semantic web technologies. With respect to building ELNs, all of the approaches discussed (UCD, ethnography and collaboration with domain experts) seem to have a high level of merit, and as such these will all be included in the building process.

Following these initial hypotheses, Chapter 3 will outline the overall methodology of this thesis, illustrating it's chronological development and explaining how the rest of the research progressed from this stage. Chapter 4 will then discuss the initial user studies that took place to gain a better understanding of the lab environment and the current lab practice, to both refine this idea and see where this type of system could potentially fit into the lab environment.

Chapter 3

Research Methodology and Design

This chapter will detail the different aspects of the methodology used in this thesis to formulate the studies that will be used as a basis for creating the proof of concept ELN. It will begin by describing the socio-technical theories that have been used to understand and interpret the results of the user studies. This will be followed by an explanation of the methodological approaches used to carry out the user studies. The chapter will then go on to describe the iterative design process for evaluating SemantiCat to answer the primary research question, with a a step by step description of the different research phases of this thesis. The detailed methodology for the user studies described in Chapters 4 and 6 will be explained in their respective chapters.

3.1 Theories

It is important to note that the Web is a socio-technical construct, and therefore needs to be considered on more than just a purely technical level, as described by both Hendler et al. (2008) and Berners-Lee et al. (2006); they detail how the Web is 'part of a wider system of human interaction', explaining the affect it has had on society, not least due to the wealth of information that it has made available to wider ranges of the population. It is imperative therefore, in an interdisciplinary endeavour such as this thesis, to research and incorporate the social perspectives as well as the technical. An interdisciplinary approach is key, not just between chemistry and computer science, but between the sociological and technical approaches.

In a general sense, social theories are paradigms or frameworks that seek to explain the mechanics and interplay of social phenomena (Hedström and Swedberg, 1998). With respect to the user studies conducted in this thesis, two socio-technical theories that are relevant to the web and the development of information systems will be looked at and explained, with details as to how they will be used in the research conducted. The UTAUT that looks at the factors required for adopting a new technology, and Actor Network Theory (ANT) which looks at both human and non-human actors in a network will be considered in relation to this thesis.

3.1.1 Actor Network Theory (ANT)

This theory defines 'actors' as human and non-human entities within a network, and emphasises the importance of the roles of 'heterogeneous' actors, within such a network (Latour, 2012). Previous approaches to researching technical innovation have produced some stark contrasts. Technological determinism assumes that the adoption of any technical innovation can be solely attributed to its technical affordances, and that technology drives the development of social structure and cultural values, rather than being influenced by them (Woolgar, 1997). Social determinism takes a position of the other extreme; whereby it gives a higher credence to the social over the technical with respect to technology adoption and can overlook the role of technical affordance as exemplified in the SCOT (Social Construction of Technology) approach developed by Machina and Wild (2013), which argues that almost every stage in the development of a new technology is shaped by the social groups it intersects with.

Actor Network Theory essentially looks to refute the idea that for a technological innovation to be accepted and adopted (in other words for a new social or technical actor to be accepted into an existing network) depends on either purely the social or the technical; any technological change must stem from both the social and the technical (Tatnall and Gilding, 2005). In order to give credence to both human and non-human actors, three principles of ANT have been developed (Latour, 2005):

- **Agnosticism** - impartiality towards all actors.
- **Generalised Symmetry** - explaining the viewpoints of different types of actors through a neutral vocabulary that has been constructed to work the same for both human and non-human actors.
- **Free Association** - both eradicating and abandoning any predisposed distinctions between the social and the technical.

The importance of all actors and the ever-evolving nature of networks is also emphasised, and this theory aims to achieve neutrality towards all actors through these principles. This is a relevant theory to this thesis, as to investigate current lab environments, and to attempt to bring the modern power of the web to chemical research, both the technical components and the people involved at every stage of this cycle are equally important; this thesis therefore studies both the technology and the users.

Whilst there are no obvious occurrences in previous literature of ANT being applied to electronic lab notebooks, it has been applied to similar studies of 'disrupting' a network by adding new technological components to a traditionally paper based system within the domain of healthcare. Greenhalgh and Stones (2010) uses the principles of ANT to explain the struggles that actors face with respect to electronic record technologies; ANT will be used in a similar capacity in this thesis. In traditional paper-based systems, the scientists are actors in this network as is their 'paper lab book' and their interactions must be studied, both from the social and the technical perspective. The lab notebook can be considered as an actor because it is a vital component of the lab process; facilitating experiment coordination, planning and note taking. The comments and reactions from scientists regarding Paper Lab Notebooks (both in the previous studies detailed in this Chapter, and in later Chapters from the user studies conducted as part of this thesis) emphasise its importance in the lab process, and how removing it would disrupt their practices.

The role of the paper notebook therefore needs to be examined in relation to how it shapes the social interactions and behaviour of the scientists. This also holds true for the concept of the Electronic Lab Notebook, which can be looked at as a new actor that is being introduced into the network of the laboratory. ANT hypothesises that if an actor is added or removed from the network, then it will have an effect on the entire network (Tatnall and Gilding, 2005); meaning that the 'effect' the ELNs has had (in the instances where scientists have used ELNs in the past) or is perceived to have on the existing network of the scientists in the lab environment must be considered. Additionally, the potential effects and attitude to the removal of the Paper Lab Notebook must also be considered, as this can equally be considered as an actor within the lab environment. There has clearly been resistance to the idea of removing the PLN, therefore its presence as an actor must be important, and it would obviously disrupt the network of the lab. Therefore, the network of the lab needs to be studied an overall working system, and this theory will be used to better understand the disruption that would be caused to current working practices by removing (or replacing) the Paper Lab Notebook, and adding the Electronic Lab Notebook.

3.1.2 Unified Theory of Acceptance and Use of Technology (UTAUT)

As a complementary or alternative model to the socio-technical ANT, another model that is highly pertinent to this thesis is the UTAUT. This was conceptualised by Venkatesh et al. in 2003. This theory builds on former technology acceptance based theories such as the Technology Acceptance Model (TAM) (Davis, 1989), and other social and technical theories to try and explain what factors would influence a user to use a piece of technology It postulates that the four main variables that will influence the usage are:

- **Performance Expectancy** - how much an individual believes that using this technology will improve their job performance (similar to perceived usefulness in TAM).
- **Effort Expectancy** - how easy the technology is to use (similar to perceived ease of use in TAM).
- **Social Influence** - how much somebody more important than an individual thinks that they should be using this technology.
- **Facilitating Conditions** - how much an individual feels that the technology is well supported with a suitable technological and organisational infrastructure.

These are all factors that need to be considered with respect to this thesis. With respect to UTAUT it has become clear from the previous literature that ELNs have both been perceived as difficult to use and not particularly worth using in some instances (Shankar, 2004, 2007; Hughes et al., 2004a; Oleksik et al., 2014; Guerrero et al., 2016). This makes it even more important that comprehensive user studies are conducted to identify what features would be required to fulfil the technology acceptance criteria.

The previous ethnographic studies looked at in Chapter 2 highlighted that scientists have struggled with the idea of giving up paper, based on its many affordances, one of which is how easy it is to use. Therefore the features that are associated with papers ease of use (effort expectancy), and an ELNs subsequent less usable features need to be identified. If a chemist is going to adopt a new piece of technology (aka an ELN) then they need to be convinced that it is a) easy to use and b) that using it will actually provide some improvement to their current state, and in this case probably add more value than the original paper based method, or otherwise what is the incentive to make the change? Therefore, the characteristics that the scientists would perceive to be useful in an ELN also need to be determined to identify the appropriate place for an ELN to fit into the note taking process such that a user's performance expectancy could be increased.

Additionally, social influence is an important factor to consider, as figures in authority (either supervisors/mentors for academia, or bosses/managers both in academia or industry) could have an impact on whether an individual makes an effort to adopt a particular technology, or in how they perceive it in the first place. Finally, the support network behind a system is also worth consideration, as discussed earlier in this chapter, scientists find creating their records a very personal endeavour and have their own unique ways of organising them on paper exactly how they want. It therefore follows that they would need a well-supported system that facilitates recording and organising their notes, and is more stable than their current paper-based practice; which would not be possible if the software was not well organised and didn't have the appropriate technical support in place.

3.1.3 Using ANT and UTAUT

Both of these socio-technical models will be used to help evaluate and understand the results of the user studies conducted in Chapters 4 and 6. Despite coming from different subject areas these themes can be looked at in conjunction with each other, as in an ideal world the ELN would be viewed as a welcome addition to the network, that is easy to use and considered actually useful. Adopting a new technology is more than purely a technological or social process as illustrated by the UTAUT. The user studies and research done towards current lab practice and the needs of the scientists involved will aim to identify the characteristics that an ELN environment would need to fulfil this brief, and will look to better understand any social influences or wider concerns surrounding this area than just using the technology.

These models complement each other as they overarchingly consider similar concepts with respect to giving credence to both the social and the technical factors of technology adoption. Both theories however provide different useful perspectives with which to consider these socio-technical aspects. UTAUT focuses on the factors that would influence a scientist to adopt and use an ELN, both from a technological perspective of a user's perception of technology, and how external social factors influence their software usage; whereas ANT studies the interactions between the human and non-human actors (the scientists and the paper and electronic lab notebooks) to understand what impact removing the paper lab notebook and adding the electronic lab notebook to the network would have on the current workflow practices, and takes a more holistic view of the other social factors associated with using ELNs.

These theories whilst considering different perspectives can be used in conjunction with one another. ANT can be used to illustrate how the current practices would be disrupted if an electronic lab notebook was added to the lab work flow, either as a new addition or a replacement to the paper lab notebook. However, the elements of 'disruption' can be considered using UTAUT; as the adoption factors that are key to this theory help illustrate *why* an ELN would cause such disruption in the first place.

The adoption barriers detailed in Chapter 2 illustrate many reasons why scientists are resistant to using an ELN including Ease of Use (see section 2.7.1.3) which directly links to UTAUT's Effort Expectancy factor. 80% participants of the Dial-a-Molecule survey believed that certain types of information would be difficult to capture in an ELN, which is both an adoption barrier, but also an illustration of how an ELN could cause disruption to the current workflow practices, as inputting information into one as opposed to using a paper lab notebook could prove more difficult and time consuming, therefore disruptive to a scientists standard workflow. Similarly the adoption factor of performance expectancy could suggest that in order to accept a disruption to a network (e.g by adding a new actor) perhaps scientists would need to perceive that this new actor would improve their current practices, therefore making the disruption worthwhile.

Furthermore, there is also the aspect of social hierarchy in the lab to be considered. UTAUT denotes that social influence (how much an individuals superior believes that they should use a piece of technology) is a key factor with respect to technology adoption. However, it's also worth noting that given the typical social hierarchy of an academic research lab, PhD students (the participant audience of the focus groups and lab observations conducted in this thesis) have their practices and workflows defined or influenced by their supervisors. The user studies conducted in Chapter 4 illustrate that many of the PhD student participants followed practices set out by their supervisors, and that students in the same lab group operated under a common workflow. Similarly respondents of the Dial-a-Molecule survey noted that they believed that adopting an ELN only made sense if the entire department adopts it; suggesting that scientists would wish to continue to operate under a common set of practices to their peers, and also that adopting certain pieces of technology could prove to be more of a group (or network) decision as opposed to an individual consideration.

This could also suggest that if an ELN was incorporated into the common workflow then it would be better supported within the organisational infrastructure (linking back to the UTAUT factor of facilitating conditions). Therefore the principles of ANT need to be considered directly alongside the social influence adoption factors to consider how important the social hierarchical aspects are, and also what impact they have to both the workflow practices and the disruption to them that could be caused by adding an ELN to the network. Overall, these theories will be considered in conjunction with one another to evaluate the results of the initial user studies and the software evaluation focus groups. The interactions between the actors as well as the different acceptance criteria required from the chemists to make these ELNs viable will be investigated hand in hand in this thesis.

3.2 Methodological Approach

Whilst there are many different types of research, research approaches have typically been split into two main forms: Quantitative and Qualitative (Kothari, 2004). Quantitative research methods are focused towards the data that can be measured or mathematically analysed. They are normally reserved for data that can be collected through surveys, questionnaires, polls or pre-existing numerical data (Muijs, 2010). Qualitative research methods look at data from either direct contact with research subjects, or documents written by them (Patton, 2005). They encompass participant observations, interviews, and focus groups.

Both of these approaches have their merits, however neither approach alone is enough to answer the research questions set out in this thesis. Understanding the current lab practice and getting to the roots of why scientists currently still tend towards paper, lends itself well to a qualitative approach. These questions will elicit richer answers by talking to scientists to ascertain their opinions and attitudes, which is a cornerstone of the qualitative approach (Kothari, 2004). However, analysing the software usage of chemists and identifying the current status of the ELN market is more conducive to a quantitative approach. This dual need for both quantitative and qualitative approaches led to undertaking a mixed methods approach. Mixed methods have been characterised by Clark and Creswell (2011) as "those that include at least one quantitative method (designed to collect numbers) and one qualitative method (designed to collect words)". There are however a number of different types of mixed method approaches, these are described in Table 3.1.

Mixed Method	Description	Data Collection
Triangulation Design (Punch, 2009)	Locate qualitative & quantitative data on the same topic that complement each other.	Data is usually collected at the same time and collated together.
Embedded Design (Clark and Creswell, 2011)	Qualitative & quantitative data are used to support each other as one dataset isn't enough to answer the research question. One dataset may end up 'embedded' in the other.	Data can be collected simultaneously or in sequence.
Explanatory Design (Clark and Creswell, 2011)	Qualitative data is used to build on and explain the quantitative data. Typically used where the qualitative data is required to fully interpret the quantitative data.	Quantitative data is collected first, then qualitative data.
Exploratory Design (Punch, 2009)	Qualitative data is collected and then followed up with quantitative data. This is generally used when the researcher requires quantitative data but needs to gain better understanding and knowledge from a qualitative study first.	Qualitative data is collected first, then quantitative data.

Table 3.1: Mixed Methods Approaches

Mixed methods are highly appropriate for this thesis as they will allow both qualitative and quantitative research to be carried out. The gaps in previous studies, combined with the research and revised theoretical concepts detailed in Chapter 2 have led to designing the following studies:

- **Software Usage of Chemists Survey** - this will highlight the areas that chemists actually use technology, rather than just focusing on whether they currently use ELNs or not.
- **Scientist Focus Groups** - this will help establish current lab practice, and understand how scientists currently operate within the lab, organise their notes and how they feel about ELNs.

- **Lab Observations of Chemists** - to complement the focus groups and enhance the understanding of both how the different chemistry labs operate, and the needs of the different types of chemists; and also to see if the chemists actually do what they say they do.

The most appropriate methodology that fits with these particular studies is the 'embedded design' detailed in Table 3.1. These datasets will be used to support each other, as they are all needed to answer the research questions, and the data will be collected in sequence. The knowledge gained from these studies combined with the literature and market research will be invaluable in contributing to both designing the proof of concept ELN and enhancing the understanding of the current state of the lab and how scientists use their paper lab notebooks. Thus they will provide the necessary information to answer the research questions set out in Chapter 1, and to see if the features identified in Section 2.6 are actually desirable to scientists.

3.3 Research Design

Chapter 2 identified three key approaches that should be taken into account when designing ELNs: ethnographic studies, collaboration and user centred design; subsequently these were used in designing the proof of concept ELN. As part of this thesis, focus groups were conducted to characterise the current lab practice and to elicit the requirements of scientists for an ELN, and also to help iteratively design the prototype software Semantic-Cat that was created as part of this thesis. An ethnographic study was also conducted to observe chemists in the lab to also gain further understanding of the current state of the lab and to gain insights towards the chemists working environment by observing them in the lab. Previous studies have used these techniques; the interviews conducted by the Smart Tea project (schraefel et al., 2004) and Oleksik et al.'s observations of the lab aided both projects in identifying the crucial needs of the user and enabling a further understanding of how the user works from their own perspective.

The outcomes of the focus groups and lab observations were therefore used to help design the initial requirements of the proof of concept ELN. Once the first prototype was created, user centred design techniques alongside collaborations with scientists were used in order to produce new iterations of the system. (Reimer and Douglas, 2004) and (schraefel et al., 2004) illustrated in their work that using UCD and collaborating with scientists at each stage of development allowed the developers to gain valuable insights into the users' needs, and aided in creating a functional tool that captures the necessary features that the users actually want.

The initial prototype that was created was primarily a backend design that offered a range of functionality, and a trial focus group was used partially to shape how the overall focus groups should be structured and to test out the questions, to form some initial ideas about the types of responses the scientists might give. The potential users (scientists) were shown the system with some test documents, and were invited to investigate the different features of the system. The initial idea had been to use the feedback from all focus groups to shape the design of the next iteration from a front-end perspective. Part of the participants feedback was that they found it confusing to understand that they were being asked about how the functionality displayed in Semanti-Cat should look in an actual notebook system. The response to this was to create mock-ups for the main focus groups, using the feedback from the initial focus group to shape the first design as the participants had observed the potential functionality and helped conceptualise how these features could actually be added into a notebook in a way that they would actually be useful to a user.

Three more focus groups were then run with physicists, chemists and biologists to test the functionality of the system with their own documents, and to provide further feedback on both the design and feature offerings. This feedback was then used to produce a proposal for the next iteration of this proof of concept software that has been detailed in Chapter 7.

3.4 Research Analysis

To analyse the quantitative data, the totals for each question were plotted on graphs to address the specific research aims of the survey. Simple bar charts and stacked bar charts were used to interpret this data. The initial charts were created to see if there was any obvious correlation between career stage and types of software used, and types of chemists and software used. Following that, the overall usage of the different software types were modelled, and then for each different software type the software packages that had above zero entries for using the software package rarely or often were plotted on stacked bar charts.

With respect to analysing the focus group data, there are a number of different approaches that can be taken (conventional, directed, or summative), and these largely depend on the type of study that was performed and the aims behind it. As noted by Hsieh and Shannon (2005), if the questions asked to study participants are more open ended (e.g semi or unstructured interviews or focus groups) then typically this lends itself to using a conventional approach to content analysis. This is generally used to capture participants feelings towards a particular phenomena that hasn't been widely investigated and involves deriving the categories and codes from the data itself (Kondracki et al., 2002). Whereas in studies that are investigating well examined phenomena (where research and theories already exist) for further information researchers typically take a directed approach, which involves using pre-defined categories and codes (Kipping, 1996). A summative approach is used in studies where researchers are trying to understand the underlying context of a phenomena, which they do by identifying and defining the meaning of key words with the aim to better understand the context within which they are used.

As described in Chapter 2, the user studies conducted as part of this thesis cover gaps in the existing literature, and both sets of focus groups (the initial ones, and the software evaluation ones) involved questions of a semi structured nature that were designed to be open ended enough to spark discussion and debate. Therefore these studies lent themselves most to using conventional content analysis, which involves using open coding to identify categories from the data itself rather than look for specific categories within it. Both sets of focus groups were recorded, and transcribed by the author of this thesis. Following this, a mind map was created for each question, using different colours to represent the three different scientific disciplines. The key elements of the data were split out for each discipline, to see how the disciplines differed, and also what commonalities existed between them. Figure 3.1 is an example of one of these mind maps.

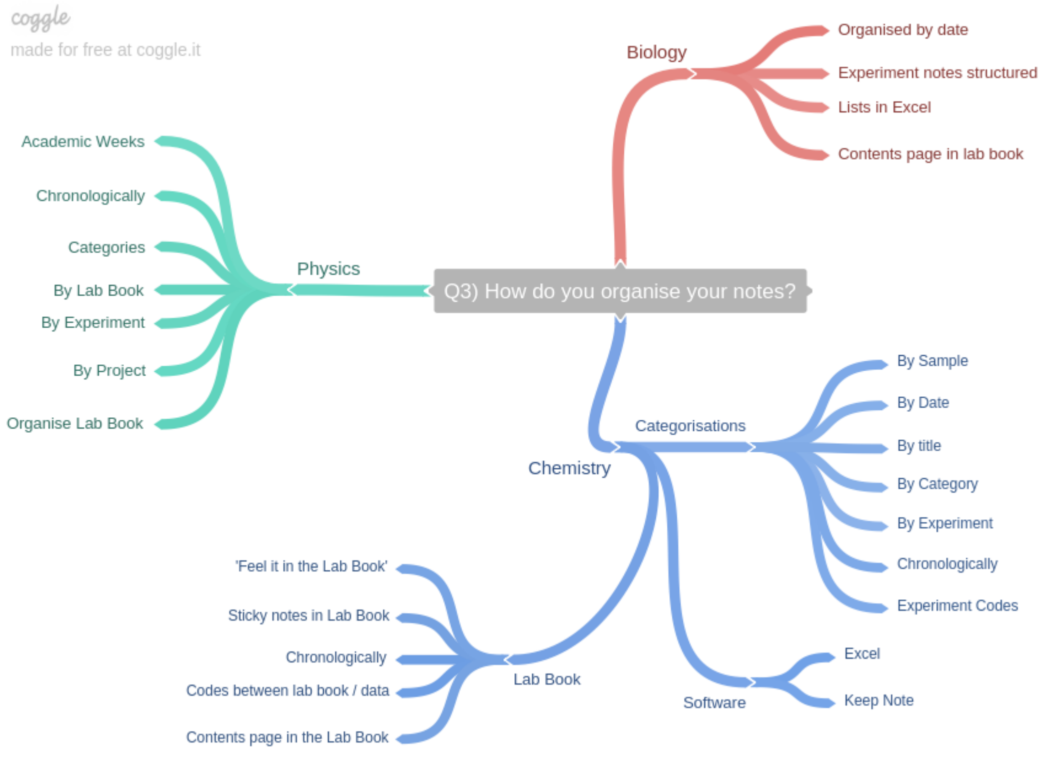


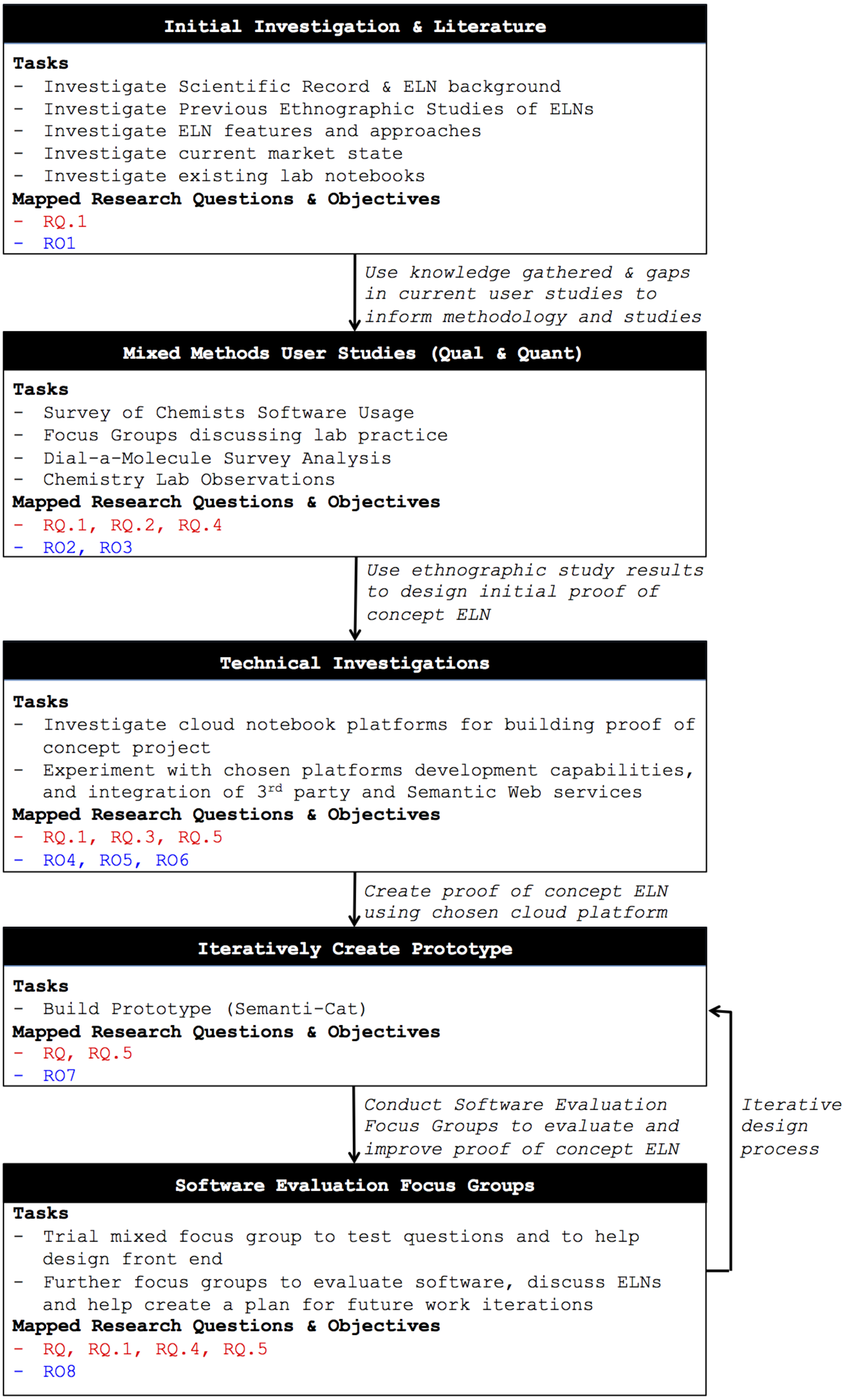
FIGURE 3.1: Mind map of the key points that came out of Question 3 from the Focus Groups, created with coggle

This mind map was created for Question 3 - How do you organise your notes. The key themes were noted down for each discipline, making it easy to see where there were common elements across the disciplines. A mind map like this was created for every question, and these were used to write up the results. Similar techniques were used to elicit common themes from the lab observations, both in relation to interpreting the results of these observations, and to understand what common themes occur across both the focus groups and lab observations together.

3.5 Research Structure

This thesis was split into five main phases, initial investigations and literature search, mixed methods user studies, technical investigations, iterative prototype creation and software evaluation focus groups. Figure 3.2 details how this research has been conducted in a sequential fashion. It shows the research tasks that have been undertaken from the beginning to the end of the thesis and highlights which research tasks/areas map to which research objectives/questions. The initial investigations and literature search has been written up in Chapter 2, and the initial user studies that formed the basis of the mixed methods user studies are discussed in Chapter 4. For each a detailed methodology is given, explaining how each study has been conducted and describing the participants involved.

Chapter 5 covers the technical investigations that were made based on the outcomes of the initial user studies combined with the conclusions drawn from the literature and market research. The approach to creating Semanti-Cat, the proof of concept software is explained, including the investigation and justification of the different platforms and services used. Chapter 6 then details the software evaluation focus groups that took place to evaluate Semanti-Cat. Again, the detailed methodology of these focus groups has been described in detail in Chapter 6 before the results and discussion. Finally, Chapter 7 will discuss the conclusions of the research questions laid out in Chapter 1, and as part of the future work section, the effort needed to create the next iteration of Semanti-Cat will be detailed.



Initial Investigation & Literature

Tasks

- Investigate Scientific Record & ELN background
- Investigate Previous Ethnographic Studies of ELNs
- Investigate ELN features and approaches
- Investigate current market state
- Investigate existing lab notebooks

Mapped Research Questions & Objectives

- RQ.1
- RO1

Use knowledge gathered & gaps in current user studies to inform methodology and studies

Mixed Methods User Studies (Qual & Quant)

Tasks

- Survey of Chemists Software Usage
- Focus Groups discussing lab practice
- Dial-a-Molecule Survey Analysis
- Chemistry Lab Observations

Mapped Research Questions & Objectives

- RQ.1, RQ.2, RQ.4
- RO2, RO3

Use ethnographic study results to design initial proof of concept ELN

Technical Investigations

Tasks

- Investigate cloud notebook platforms for building proof of concept project
- Experiment with chosen platforms development capabilities, and integration of 3rd party and Semantic Web services

Mapped Research Questions & Objectives

- RQ.1, RQ.3, RQ.5
- RO4, RO5, RO6

Create proof of concept ELN using chosen cloud platform

Iteratively Create Prototype

Tasks

- Build Prototype (Semanti-Cat)

Mapped Research Questions & Objectives

- RQ, RQ.5
- RO7

Conduct Software Evaluation Focus Groups to evaluate and improve proof of concept ELN

Iterative design process

Software Evaluation Focus Groups

Tasks

- Trial mixed focus group to test questions and to help design front end
- Further focus groups to evaluate software, discuss ELNs and help create a plan for future work iterations

Mapped Research Questions & Objectives

- RQ, RQ.1, RQ.4, RQ.5
- RO8

Figure 3.2: Research Design

Chapter 4

Initial User Studies

This section will detail the three initial user studies (and their main goals) that have been undertaken as part of this thesis, and the collaborative studies that were analysed as part of this thesis, and published alongside the work done in this thesis in (Kanza et al., 2017). The studies are detailed below in the order that they will be written about.

- Survey to Investigate the use of Software for Chemists
  - *Main goal: identify where chemists use technology.*
- Focus Groups to Examine Scientists Current Note taking Practices
  - *Main goal: establish current lab practice.*
- Dial-a-Molecule Survey Analysis
  - *Main goal: understand what features and priorities users want in an ELN.*
- Lab Observations to Characterise Current Lab Practice for Chemists
  - *Main goal: understand how different types of chemists work and what their different environments are like.*

As discussed in Chapter 2 these have been identified as areas that are lacking in the current ethnographic studies, and following the methodology outlined in Chapter 3 these studies use both quantitative and qualitative approaches. The aim behind this survey was to find out, ELNs aside, what chemists software requirements are. This is vitally important information, as in order to create domain specific add-ons to a pre-existing notebook environment, the most utilised domain specific software areas need to be identified. A vast majority of the ELN investigations and studies have looked at why scientists still use paper over ELNs and have discussed the shortcomings of certain software packages, this survey however has looked at the types of software packages chemists actually want to use. It also aims to find out if there is any correlation between age of chemist, type of chemist, and what types of software they used. This should enable the design of a system with add-ons that will be useful to chemists.

The focus groups were then conducted across the three different scientific disciplines (biology, chemistry, and physics), where all of the participants were PhD students. This is because it's important to characterise the unified and differing requirements across the different scientific disciplines. The ethnographic studies of biologists with respect to ELNs such as the works of (Reimer and Douglas, 2004) suggested that the main requirements of an ELN are synonymous with those of an EN, however, some studies of using generic note-booking software for chemists were felt to be lacking in the specific domain knowledge, detailed by (Walsh and Cho, 2012) who investigated the use of Evernote as an ELN. Newer studies such as the works of (Weibel, 2016) have also investigated using cloud based notebooks in the laboratory, but this study was focused towards undergraduate users in a specific type of lab, which still only elicits one set of requirements for that type of lab. Studying these needs and the current practices cross discipline would greatly enhance the knowledge of the different discipline needs and practices and enable the identification of both cross discipline and specific discipline requirements.

The Dial-a-Molecule group ran three surveys in 2011, an initial survey to gain knowledge and understanding about attitudes towards ELNs (these formed some of the results for the adoption barriers detailed in Section 2.7.1) and then two surveys at the beginning and end of the trial of their ELN. These surveys identified a number of features that users said they wanted from ELNs, and elicited the priorities of certain features. This data was analysed as part of this thesis, and written up in conjunction with the results of the studies conducted in this thesis.

The observations looked at the different types of chemists in their lab environments, to give an insight into what users say they do compared what they actually do, and to give a better idea of how the different chemists work. These observations were also focused towards in what instances the participants used their lab books and computers, and a further in-depth discussion about why they are doing what they are doing and why they have made the choices to use paper or technology. The lab observations conducted in the previous ethnographic studies detailed in this project have typically involved spending several months in the same laboratory, which allows for an in-depth insight of one specific laboratory, but doesn't allow for comparisons between different strands of scientists; which is one of the aims of these observations. These have therefore chosen to look at several different chemistry labs in a shorter space of time. The labs observed were all Postgraduate labs in the Chemistry department at the University of Southampton which were frequented predominantly by PhD students with some post docs.

The individual methodological approaches to these studies will be detailed, followed by the results and subsequent discussion of these studies. In the results section, for the survey, as this covered scientists at a range of career stages, the participants will either be referred to as participants, scientists or by their type of chemistry if talking about a specific group (e.g organic chemists). For the focus groups and lab observations, similarly the participants of the studies will be referred to either as PhD students or participants, but when referring to a specific set of them by discipline then that disciplinary term will be used (e.g chemists, physicists, and biologists).

4.1 Survey Design

Research has suggested that the questionnaire based survey is the most popular of the quantitative research methods (Pearce et al., 1966) due to the nature of the wide scope of information that can be collected from it, in addition to being 'logical' and 'specific' (Babbie, 1973). The main disadvantage of surveys is a potential unwillingness to answer them which can lead to 'non-response errors'; or participants might target their answers towards the researchers goals (Hart, 1987).

Surveys can either be online or paper based, however conducting them online can be advantageous (Wright, 2005). They are easy to create with specifically designed tools such as iSurvey[1] and the wide reach of the Web makes it possible to reach groups of people who are physically far away, and also reduces the cost of producing paper copies of the survey (Wright, 2005). This doesn't guarantee that everyone can be reached via the Web, or that people won't find online invitations too irritating to respond to, but in this case there are potential advantages of reaching a far wider audience than the PhD students in the University of Southampton.

In order to create the survey, different types of chemistry software was searched for using Google initially and several sites that had compiled lists of different types of software were identified and these results were combined to list all the currently active software packages (which totalled 242 unique software packages). These sources are listed in Appendix B, including details regarding the ethics applications and links to the full supplementary data. This vast number of software tools was considered too large to formulate into a question that merely asked participants which software packages they had used, both from the perspective of a survey that wouldn't be particularly user friendly, and which may not give helpful data.

Therefore these software packages were broken down into different types of software. Some of these software lists that were used to identify the different software offerings already used certain categories, which were considered in the categorisation process, alongside consulting Professor Jeremy Frey at the University of Southampton for chemistry expertise in order to accurately identify where different software packages belonged. From these processes, nine categories were devised for this survey: Molecular Modelling & Simulation Software, Molecular Editing Software, Quantum Chemistry Software, Organic Synthesis Software, Nanostructure Modelling Software, Chemical Kinetics & Process Simulator Software, Chemical Database & Informatics Software, Chemistry bibliographic Databases, Computer Based Chemical Terminology Software (Semantic Web)

[1] `https://www.isurvey.soton.ac.uk/`

and Other Software. Most of the tools identified fitted neatly into a single category but there were a few such as Avogadro[2], which fitted into both Molecular Editing and Semantic Web Software.

The survey begins by asking what type of chemist the participant are, and how many years' experience they have. The types of chemist are: Analytical, Environmental, Industrial, Inorganic, Material, Organic and Physical. This was to see if there was any correlation between a chemists level of experience, and what types of software (or how much software) they used to see if there are any specific groups that either don't use software or heavily use software. The chemists were asked about the type of chemist they were to better understand what types of chemists use different types of software. Chemistry is a rich science with many sub disciplines within it, and in order to produce software that could cater to this level of diversity within a discipline, it is important to understand which types of software different chemists need.

The survey is then split into 11 further sections, 10 of which ask about the different types of chemistry software detailed above, first to ascertain if they've used it or not, and if so to ask if they've used any of the listed software packages that fall into those categories. This was both to see which types of software were most used, and if there were any obviously popular packages that a majority of chemists used. This would enable identification of not only the types of software that should be included as a domain based add-on in an ELN, but would also point towards well used software that could be examined either to see if there was an open source version that could be used, or what types of functionality were included in it.

The final question asked was what tool the participant would most like to be created. These questions were designed not only to get a better feel for what types of tools are used by chemists, but also to be able to correlate which types of chemists use which types of tools. As there are also so many different types of tools, this will also demonstrate how many are actually widely used. This should build up a better idea of what type of add-ons would be appropriate for a chemistry ELN environment, as popular open source tools could be converted into add-ons for an ELN environment. In order to design an ELN that would meet with a scientist's domain based needs, it is important to understand what types of software are required.

These questions were specifically designed to meet the Research Objective RO2, to ascertain what types of software chemists use, to help answer the Research Question RQ.1: What are the approaches and features that should be taken into account when creating an ELN? The research conducted in Chapter 2 suggested that ELNs would require domain specific knowledge and features, especially for chemists. Therefore, in order to understand what types of features should be included, this survey aimed to find out what areas chemists used software; and for each software category, the most popular

[2] `http://avogadro.cc/wiki/Main_Page`

package that could feasibly be included as a domain based add-on for an ELN was considered. This information was combined with the focus groups and lab observations to help explain why this software is used, after the survey has identified what has been used.

4.2 Focus Groups

Focus Groups are a form of group interview (Kitzinger, 1995) that also takes advantage of the additional communication between the participants. Interviews facilitate a back and forth conversation exchange between the interviewer and interviewee, whereas a focus group can take advantage of the extra conversation or debates sparked by comments from certain participants (Kitzinger, 1995).

The advantages of this method are that it has the capacity to generate a lot more data; participants may give more information through informal methods of communications such as responding to another group member or making a joke. It also has the potential to allow people who are more shy and reticent to participate, as they might be more willing to either join in as part of a group, or to talk more once others have initiated a discussion. The debates and conversation that can occur as part of a focus group can also provide new perspectives that enable participants to share their thoughts because they find an element of the discussion they can join in with (Kitzinger, 1995).

Alternatively however, a disadvantage of focus groups can be that some groups will have louder more dominant individuals who might intimidate or talk over quieter members of the group. Additionally, the more participants are involved, the more it can impact on confidentiality as every participant is privy to what the others have said. King (1994) notes that people like to talk about their work, but aren't often presented with the opportunity to do so, which might encourage them to take part in such studies.

These focus groups were designed to use a semi structured set of questions, to lead the discussion to certain topics whilst still leaving it open enough to facilitate discussion and debate (Britten, 1995). The participants who were approached to take part in these focus groups were PhD students studying either chemistry, physics or biology. Despite this thesis having a chemistry focus, as noted in Chapter 2 many of the previous ethnographic studies tended to focus on one singular scientific discipline; therefore the three main scientific disciplines were all included in these focus groups to help understand what ELN features all scientists needed irrespective of discipline, and to elicit the different needs that exist for each discipline.

The focus groups were run after the survey, and were used to collect data about what the participants think that they do, and facilitated discussions between them about their current note taking practices and how they organise their data, as well as ascertaining their opinions on ELNs. Arguably these first two studies could have been run simultaneously or have had their order reversed, but for the purposes of planning and ethics applications they were done in this order.

The questions begun by asking what methods the participants used to record their notes, to understand whether they used paper, electronic devices, or a combination. They were then asked about how they record different types of notes (e.g notes, mind maps, graphs, pictures, photos, diagrams, tables) for different activities which were: conducting experiments inside and outside the lab, looking at literature, thinking about work, performing calculations and writing up work. This was asked to gain a better understanding about which tasks the PhD students felt lent themselves better to paper or electronic note taking, and how they approached different types of work. It was also asked to facilitate discussions about what technology was used if any to aid with different pieces of work.

In a similar vein, the participants were asked about their use of technology in note taking to understand where they already used technology and to elicit any inadvertent uses that they may have not considered such as emailing themselves documents. This was asked with similar motivations to the survey in order to understand where the participants already used technology. The participants were also asked about how they organised their notes, and how they linked together their paper and electronic notes to see what practices had been put into place and how much different PhD student organisational patterns differed. Shankar's 2007 study concluded that writing up one's notes was a very personal endeavour; this question was asked to see if organising notes was just as personal to scientists. Additionally, rich semantics had been identified in Chapter 2 as a feature to include in an ELN, therefore asking about how the participants currently link together their data would give valuable insights into how semantic web technologies could be used to link together notes in a way that the PhD students would find useful.

Participants were also asked where they stored their data and whether they were concerned about Intellectual Property and if they had notes that specifically needed to be kept secure or that had limits on sharing. Given that this project was looking to create an ELN on top of an existing cloud service, and that cloud based had also been identified as a key feature of an ELN, it was important to understand how big a concern these issues were to the participants, as they wouldn't be willing to store their data in a place that they didn't feel was secure. With regards to sharing work, the participants were asked about their collaborations, whether they were involved in collaborative activities and whether they found them useful. This was asked due to the identification of 'collaborative' as an important ELN feature, to see if the participants regularly collaborated with their colleagues and whether they thought it was a useful endeavour.

They were then given three scenarios to discuss, regarding how they would locate a piece of work from 6 months ago, what they would do if there was a fire in their lab and all their paper lab notebooks were destroyed, and if they were indisposed for a while how would their supervisors/peers access their work if necessary. The first scenario was designed to understand how they currently searched through their work, searching is one of the key processes that electronic devices can perform substantially faster and more accurately than with paper, and improved search is also a key part of using semantic web technologies, therefore this process was necessary to understand. The second scenario was designed to understand how much the PhD students actually value their paper lab notebooks and whether they would be concerned to lose the content in them or whether they feel that it is only useful in the moment. Additionally, encouraging the participants to discuss how they valued their paper lab notebooks, and what material they felt that they would stand to lose from it, was facilitated to help identify which pieces of work do not get digitised in some form.

The final scenario was constructed in order to understand how accessible the participants made their work to others from an alternative perspective. Preserving the scientific record is important for knowledge preservation, but there is little point putting in effort to digitise and organise all of their work if it's not made accessible to anyone else. To further understand the current preservations of the scientific record the participants were asked about how they back up their electronic and paper notes as this would also clarify which pieces of work they deemed important enough to back up and which pieces of information could potentially be lost.

Finally, the participants were asked if they had ever used an ELN, what they would expect from one and how they thought it could improve their work. The literature search conducted as part of Chapter 2 suggested that the uptake of ELNs in academia has been limited, so this question was also asked to see whether these participants experiences matched with that hypothesis, in addition to facilitating discussions regarding their opinions of them and their expectations towards them.

Lastly, they were also asked what equipment they could take into the lab, as a barrier to ELN adoption that was identified in Chapter 2 as ELN Access, due to scientists concerns that there wouldn't be access to appropriate hardware in the lab for using ELNs. The full question list can be found in Appendix C. These questions were specifically designed to meet the Research Objective RO3 to characterise current lab notebook practice, in order to help answer the Research Questions RQ.1, RQ.2 and RQ.4. What approaches and features that should be taken into account when creating an ELN? What are the key processes of digitising scientific research? Where could an ELN fit into the current lab practice where it would actually get used?).

4.3 Participant Observation

There is often a disparity between what people say that they do and what they actually do, and participant observation is a method of not only observing people's actions, but uncovering these potential imbalances (Mack et al., 2005). These observations facilitate capturing the different perspectives of the participants and allow the observer to gain a better understanding of what their natural environment is like. The process of a participant observation is for the observer to make objective notes about what they have observed, as well as engaging in informal conversation and interactions with the participants (Mack et al., 2005).

This method holds many advantages as noted by Bernard and Gravlee (2014). It enhances the quality of the data collected because it highlights what the participants actually do and can uncover unknown factors that are important but were unknown during the design of the study (Mack et al., 2005). This method can also improve the quality of the data analysis (Bernard and Gravlee, 2014) as it can lead to a better understanding of data collected through other methods (Mack et al., 2005) such as focus groups or other quantitative methods; and can help design questions for future studies.

A disadvantage to this method is that it is very time consuming. Mack et al. (2005) suggests that usually observers would spend up to a year in the field, which is often not feasible for certain projects. These observations also require a level of cultural awareness of the surroundings, particularly as even with that awareness, documenting the data and reporting objectively is a difficult matter. Additionally, there is always the risk of the Hawthorne effect whereby the participants may potentially change their behaviour when they know that they are being observed (Wickström and Bendix, 2000).

Wolfinger (2002) has noted that there a number of different strategies for writing field notes in such observational situations. The chosen method for this project was to systematically note down the things that were happening temporally from start to finish of the observation. This was so that it would allow for an observation of what occurs in the lab over the course of the time spent there to see how the participants act in their natural environment ,and at what times during experiments, or times at their desks that they take notes, use their lab book, or use technology.

The participant observations were the final study that took place. This was the most useful place for them because the preliminary investigation of the survey had already been performed, and the focus groups had also been run; which provided valuable extra knowledge and insight for both observing the lab environment, and for what questions to ask during the observations. Temporal notes were taken, with detailed observations from the start to end of the observation time in the lab. Multiple different chemistry labs within the University of Southampton were approached to see if they would be willing to permit short term observations of their labs and working environments. The participants were also asked questions; the level of discussion with said participants was dependent on their level of engagement throughout the study. These observations were planned and ran after the first two user studies, in order to help meet the Research Objective RO3 (alongside the focus groups above) to help answer Research Questions RQ.1, RQ.2 and RQ.4.

4.4 Results

This section will detail the results of the survey, focus groups and participant observations. These results will then be discussed together with relation to ANT and UTAUT.

4.4.1 Survey Results

The survey was conducted first and had a total of 132 participants take part in it. Links to the raw data, the full survey, and additional data tables can be found in Appendix B. The first question in the survey asked the participants to select all the types of chemistry that they felt applied to them. Table 4.1 shows the statistics of the different types of chemists that took this survey.

Type of Chemist	% Answered	Count
Analytical	15%	20
Environmental	6%	8
Industrial	5%	7
Inorganic	8%	11
Material	14%	19
Organic	17%	23
Physical	85%	112

Table 4.1: Percentage of Different Types of Chemists in the Survey Participants

This table shows that multiple participants have chosen more than one type of chemist to apply to the work that they do, and that a majority of the participants think of themselves, partially at least, as physical chemists. This is worth considering with respect to the popular types of software as they could be biased towards physical chemistry tools given that a majority of the participants appear to work in that field of chemistry. It's also worth noting that some types of chemists are somewhat unrepresented by the participants of this survey; there are less than 10% of the participants who work in the areas of inorganic, environmental or industrial chemistry, which could also lead to a bias against the tools specifically targeted to those areas of chemistry.

4.4.1.1 Comparing Usage of Different Types of Chemistry Software

As discussed in Section 4.1, the different types of chemistry software had been broken down into 10 categories, Figure 4.1 shows the percentage of usage across these.

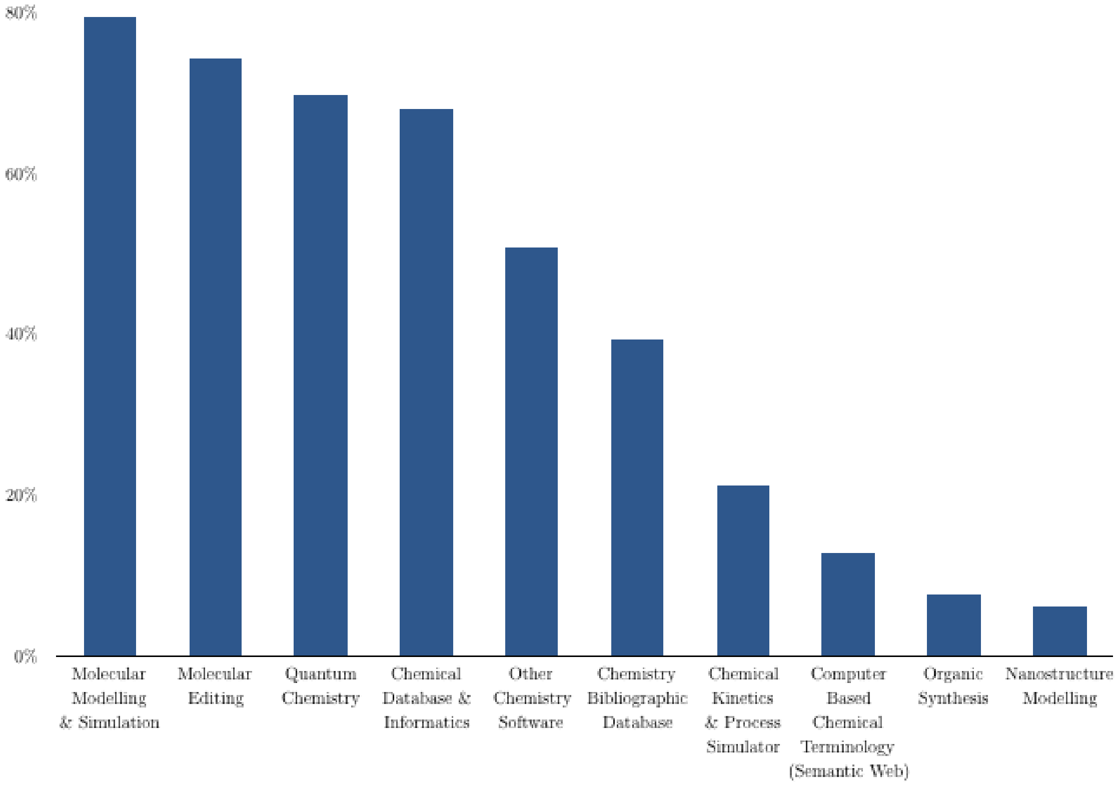


FIGURE 4.1: Chemistry Software Usage

This first graph shows that the most popular type of software used among the sample of chemists is Molecular Modelling Software, closely followed by Molecular Editing Software and Quantum Chemistry. These are all areas of chemistry that could come under the sub-discipline of Physical Chemistry, therefore given the high level of Physical Chemistry participants who took part in this survey, it is unsurprising that these types of software have been identified as the most popular. Database & Informatics software packages are also well represented here. This also tells us that despite the adoption barriers to ELNs that have been discussed in Chapter 2, chemists still have a high requirement for technology and software packages in general; which is why these software needs are being considered in the development of the ELN environment.

4.4.1.2 Comparing Type of Chemist to Software Used

This graph illustrates types of software packages used by different types of chemists. The participants were asked to tick each type of chemist that they felt applied to them, in order to to get a wider range of data, This matrix has been compiled from the data to see if there is a direct correlation between type of chemist and software usage.

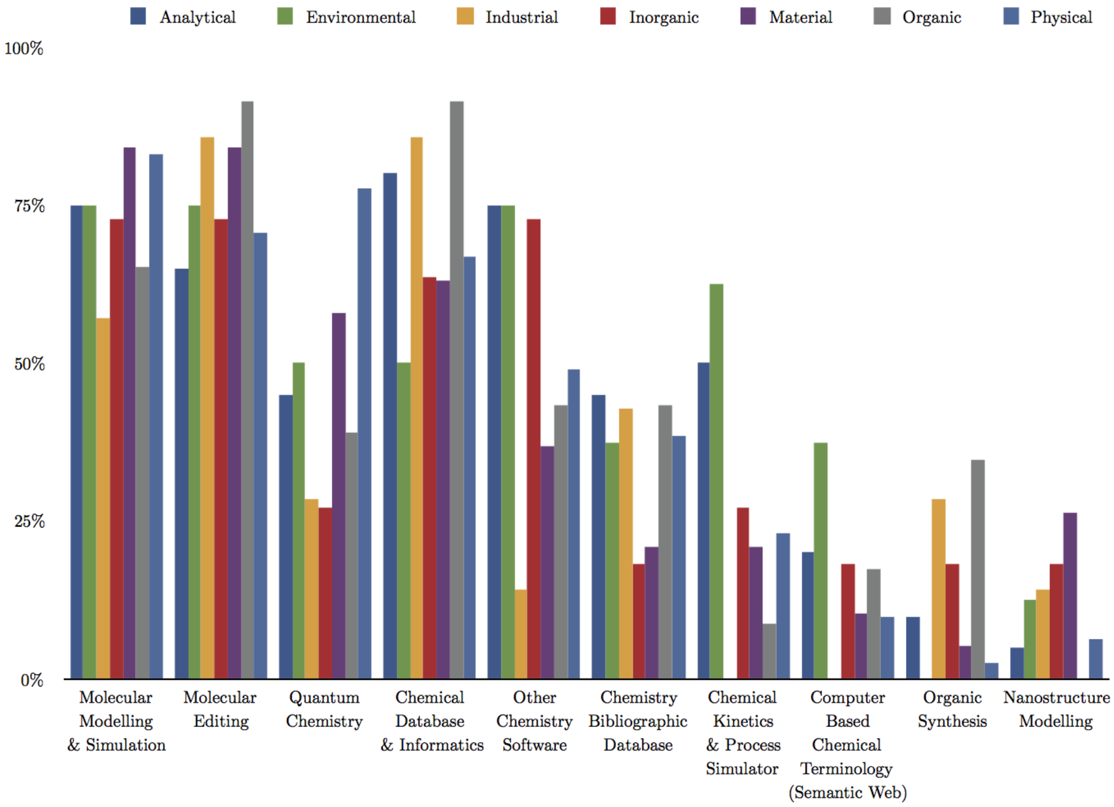


Figure 4.2: Type of Chemist Vs Type of Software Used

This graph follows a similar pattern to Figure 4.1 in terms of showing the popularity of the different types of chemistry software. There are some arguably expected results such as the highest use of organic synthesis software is by organic chemists, with the other types of chemist seemingly not using it very much, with no usage at all by environmental chemists. Alternatively, nanostructure and modelling software also shows no usage by organic chemists, with the highest usage by physical chemists. Molecular modelling and molecular editing both show a high percentage of use by most of the different types of chemists. Molecular modelling software seems to be the most popular between physical and material chemists (perhaps suggesting why this is the most popular piece of software overall, as 85% of the participants identified themselves as physical chemists), and molecular editing software seems to be most used by organic chemists. Quantum chemistry has proved far more popular with physical chemists, whereas chemical kinetics and process simulator software is most used by more environmental chemists and not used at all by organic chemists. The databases show less usage by inorganic and material chemists, and other types of software outside these categories shows a much lower usage by industrial chemists. Semantic Web software doesn't show a high level of use by any type of chemist and absolutely none by industrial chemists. This does indicate that certain types of chemists use some types of software more than others, and that there are some types of software that only get used by specific groups, but that overall certain types are more popular on an overall basis such as molecular modelling and editing software.

4.4.1.3 Comparing Number of Years Experience vs Software Usage

Another point to consider is whether the number of years' experience as a chemist makes any difference to the types of software used, Figure 4.3 shows the proportion of software usage plotted against the number of years' experience that each participant has.

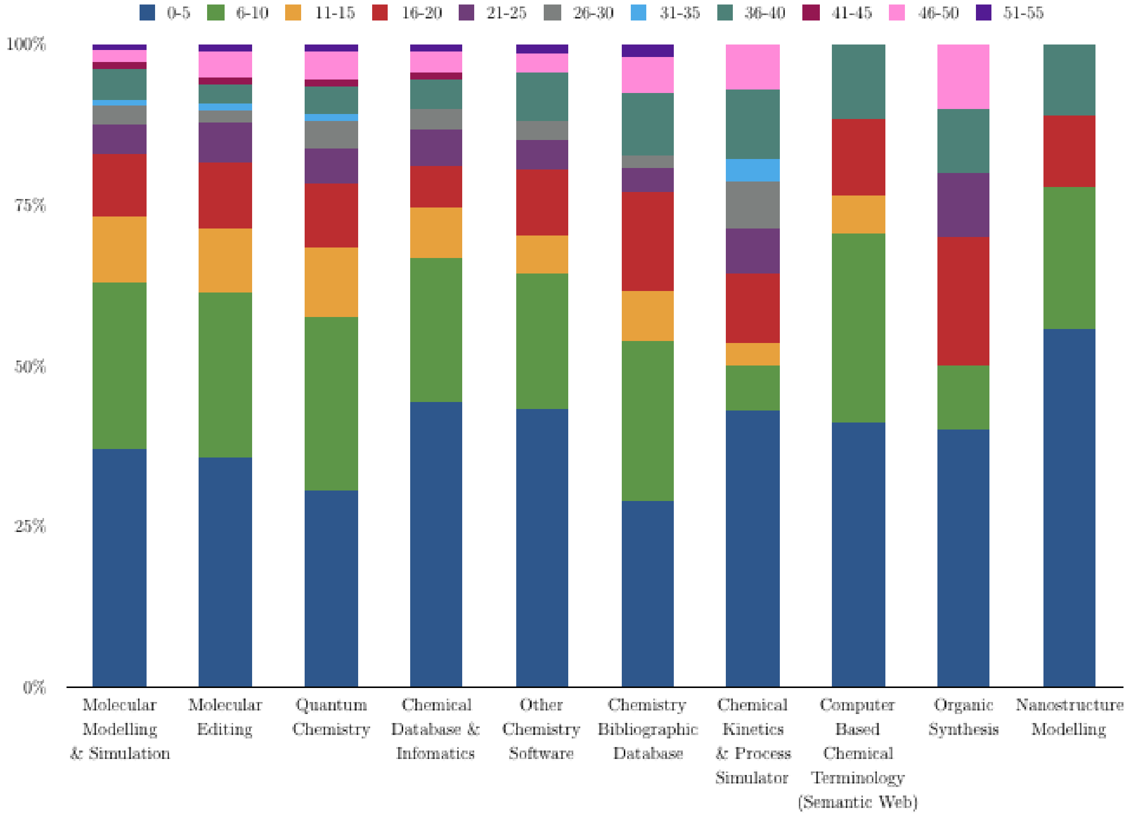


FIGURE 4.3: Years' Experience Vs Type of Software Used

This graph shows that all of the different types of software have been used by people with one to five years' experience, suggesting that there doesn't necessarily need to be a certain level of experience to use different types of chemistry software; which is interesting as it challenges a stereotype that more experienced scientists who were very set in their ways might have been slower to take up new software. The areas of software such as organic synthesis that show less usage across the different experience groups are the areas that showed the lowest levels of usage overall, so this is unsurprising. The next set of graphs will show the usage of the different software packages among the participants, highlighting which packages are the most popular and identifying the most feasible package to be used as a domain based ELN add-on. For each graph, the unused software packages will be removed; links to the full results are available in Appendix B.

4.4.1.4 Molecular Modelling & Simulation Software Usage

This graph shows the percentage of usage of the 71 different molecular modelling & simulation software packages across the 105 participants who said that they had used this type of software. This graph shows the 54 packages that participants reported using; the other 17 packages in this category are not included in this graph as none of the participants had used them.

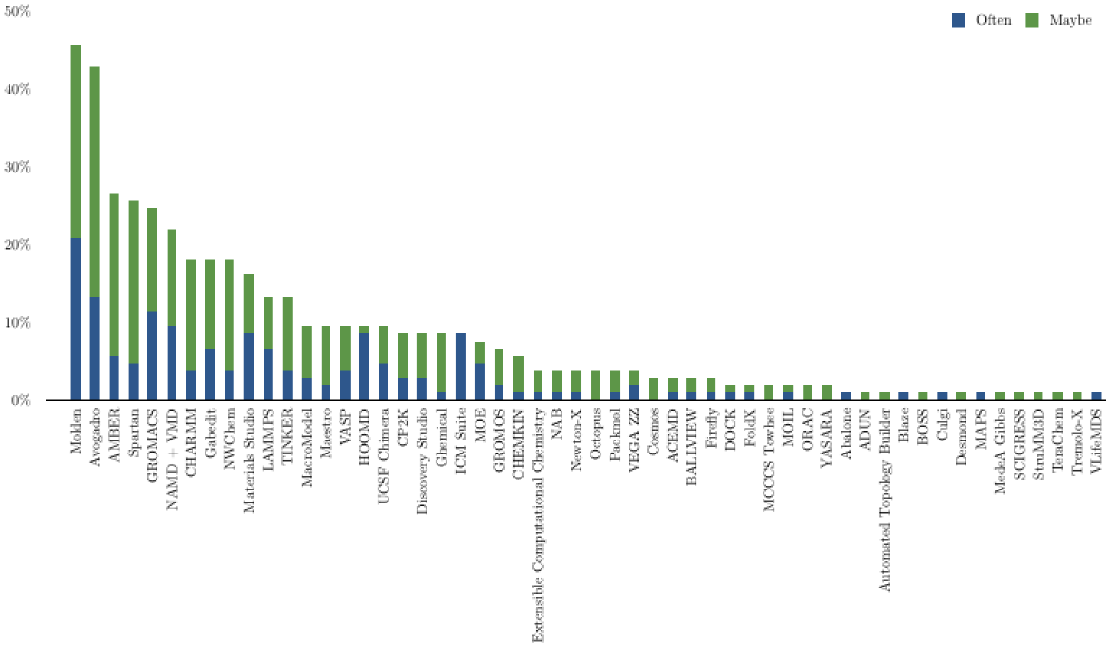


FIGURE 4.4: Molecular Modelling & Simulation Software Usage

This demonstrates that there are a wide array of molecular modelling & simulation packages available, but that not many of them are widely used, with 17 of the listed packages completely unused by the survey participants. The two obviously popular packages in this category are: Avogadro [3] and Molden[4], not only overall, but for frequent use. Molden is described as a general-purpose molecular and electronic structure processing program, so this could be because the well-used software packages are general purpose software offerings, and the rarely used programs are specialist tools. Additionally, both Molden and Avogadro are available cross platform, with Avogadro being Open Source and Molden being free for academic use, which could also have contributed to their popularity.

With regards to identifying the feasible tool for an ELN add-on, Avogadro seems the most suitable. It is both Open Source, and extensible for developers as it contains powerful plugins and the code is freely available to use (Hanwell et al., 2012), and is also one of the most popular tools in this survey's most popular category of chemistry software.

[3] `http://avogadro.cc/wiki/Main_Page`

[4] `http://www.cmbi.ru.nl/molden/`

4.4.1.5 Molecular Editing Software Usage

This graph shows the percentage of usage of the 46 different molecular editing software packages across the 98 participants who said they used this type of software. This graph shows the 30 packages that participants reported using; the other 16 packages in this category are not included in this graph as none of the participants had used them.

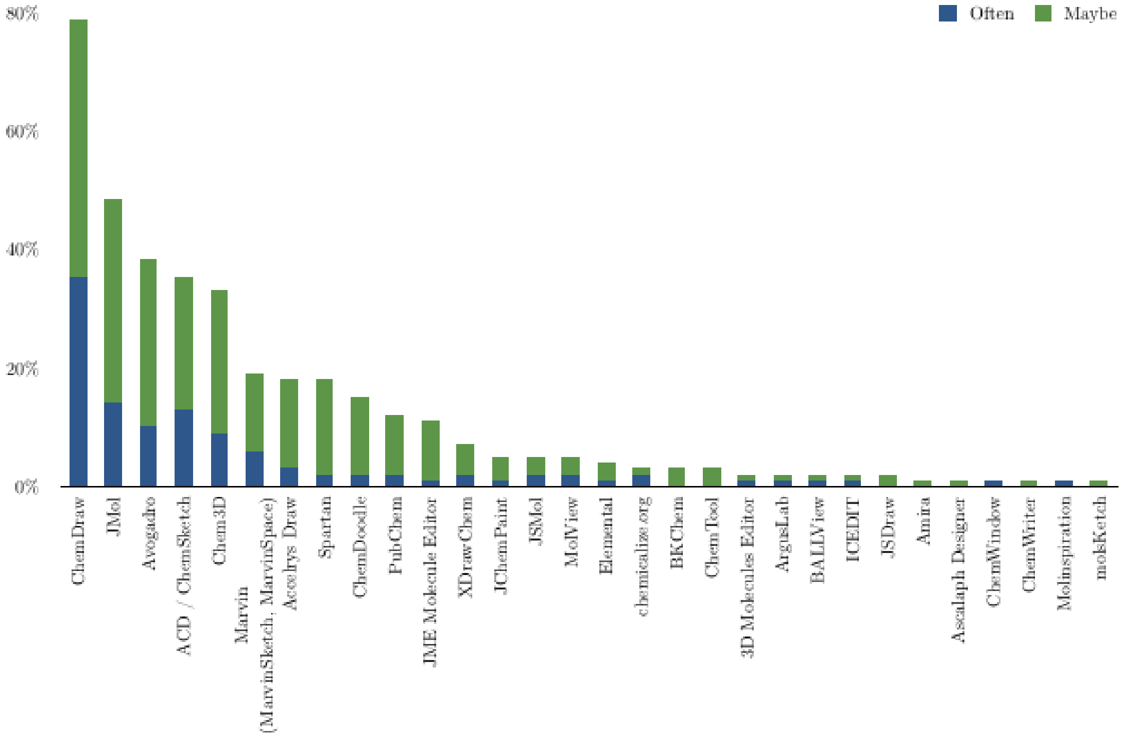


FIGURE 4.5: Molecular Editing Software Usage

Despite the survey results showing, that among the survey participants molecular editing software was nearly as popular as molecular modelling & simulation software, most of the packages listed in this survey do not seem to be widely used, with 16 of the listed packages completely unused by the survey participants. The most used by far is ChemDraw[5], and indeed ChemDraw is mentioned in both sets of focus groups (the initial focus groups, and software evaluation focus groups) as a software tool that some of the chemist participants use heavily. This is a commercially licensed software product, but it is available cross platform and on the web, and not only has free trial versions but is also significantly cheaper for academics. It is possible therefore that a majority of scientists (particularly those in academia or who have licenses paid for by their employers) use ChemDraw, and that those who can't afford it have looked for alternative offerings. These results do however show that this is clearly an area of software that people use on and off, even if there aren't that many packages that are used regularly by a large group of people.

[5] `https://www.cambridgesoft.com/Ensemble_for_Chemistry/ChemDraw/ChemDrawPrime/Default.aspx`

Based on its popularity, ChemDraw seems the best option for an ELN add-on; however this is infeasible due to its commercial nature. However, Avogadro was listed as a the second most popular piece of software in the Molecular Modelling section, and it is the third most popular piece of software in this category. This could be used as an initial tool that would also provide molecular editing functionality, and the different features that ChemDraw and Avogrardo will need to be examined in more detail to see how many of them overlap and what functionality ChemDraw provides that Avogadro doesn't.

4.4.1.6 Quantum Chemistry Software Usage

This graph shows the percentage of usage of the 75 different quantum chemistry software packages across the 92 participants who said that they had used this type of software. This graph shows the 48 packages that participants reported using; the other 25 packages in this category are not included in this graph as none of the participants had used them.

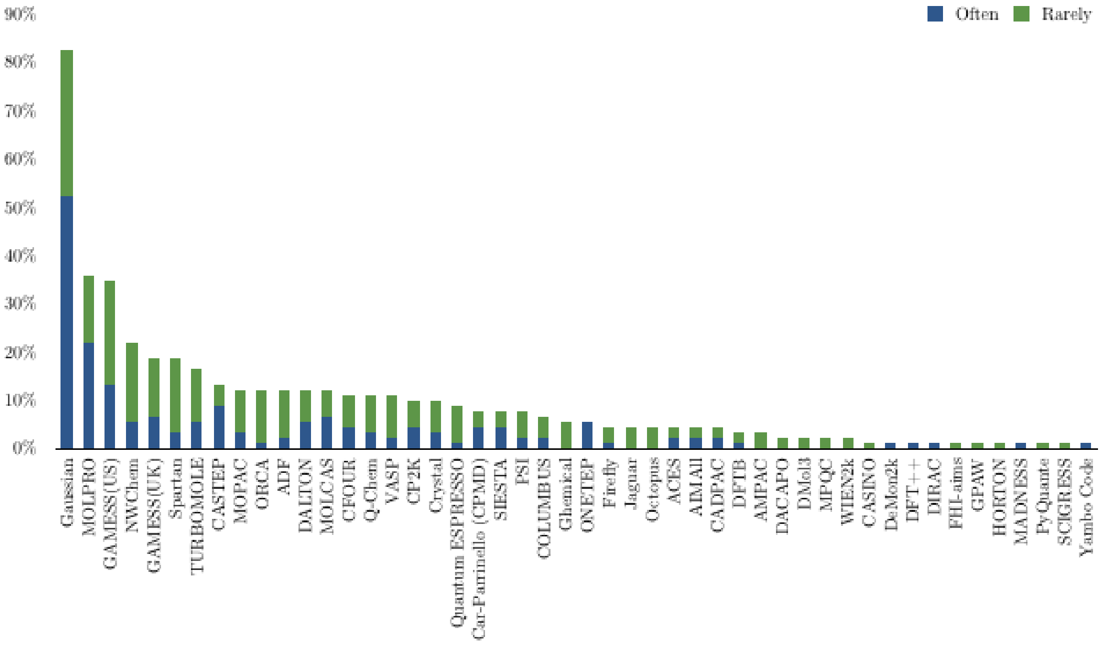


Figure 4.6: Quantum Chemistry Software Usage

This is another software area that a large percentage of the participants of this survey reported using, if infrequently for the most part. The one software package that is reported as being used significantly more often by significantly more participants than the rest is Gaussian[6]. This is another piece of software that is available cross platform, although it is a commercially licensed product, (as is the second most popular offering MOLPRO[7]) which could mean that despite its popularity not everyone can afford to use it or would be willing to pay to use it, and have therefore looked for other offering that

[6] www.guassian.com/g_prod/g09.htm

[7] https://www.molpro.net/

are more affordable (such as GAMESS[8] which is free for both industrial and academic licenses). Cost was identified as a substantial barrier in ELN adoption, which suggests that some scientists could be unwilling to pay for other software based solutions.

With regards to the most suitable software package for a quantum chemistry ELN add-on, GAMESS is the most popular free software, but its source code is not licensed for editing. The second most popular software offering in this category is NWChem[9] which is Open Source, so this would be the most viable to initially try adapting as an add-on.

4.4.1.7 Database Informatics, Bibliographic Database & Semantic Web Software Usage

This graph shows the usage of the 25 different database informatics, bibliographic database and Semantic Web software packages across the 89, 52 and 17 participants (respectively) who reported using them. These were combined due to their smaller size and similar nature, as they all facilitate accessing information and chemical data.

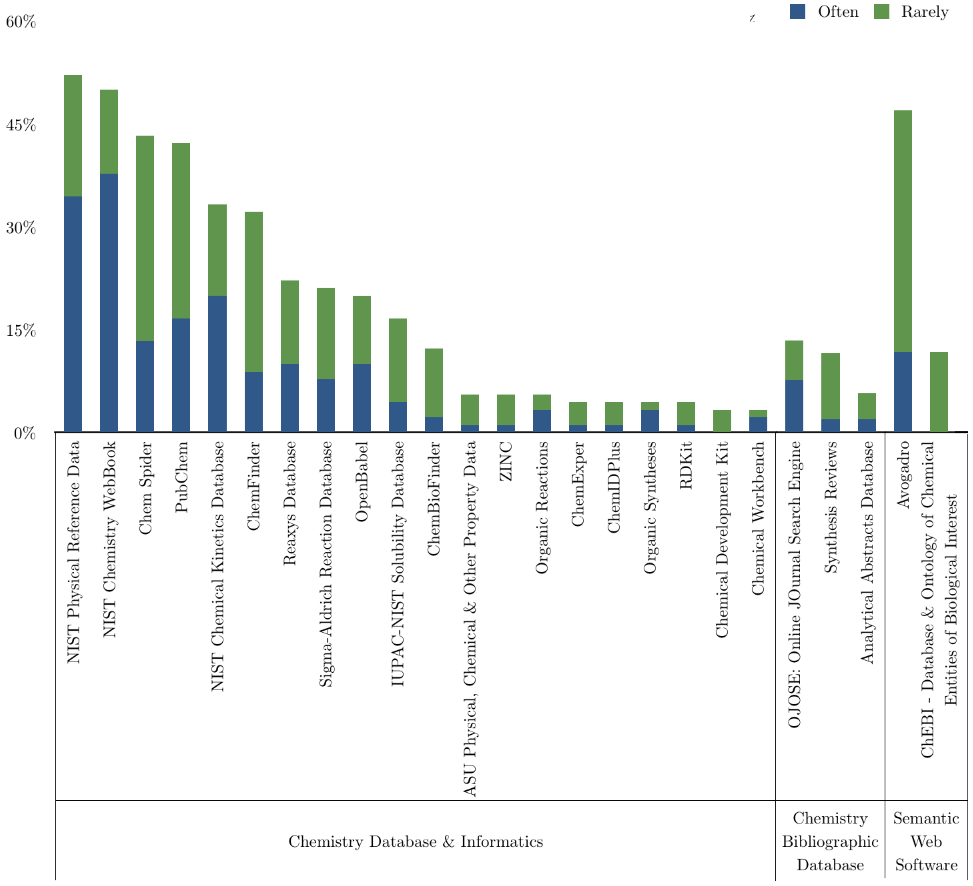


FIGURE 4.7: Database Informatics, Bibliographic Database & Semantic Web Software Usage

[8] http://www.msg.ameslab.gov/gamess/

[9] http://www.nwchem-sw.org/index.php/Main_Page

These results show that up to 50% of the participants use these types of software, suggesting that this could be an area that would be very useful for some add-ons. The most popular database and informatics software are ChemSpider[10], NIST Chemistry WebBook[11] and NIST Physical Reference Data[12]. ChemSpider has an API that allows developers to use plugins for it which is something that could be potentially integrated into the new system. A majority of these are web-based tools that allow you to look up large amounts of information, which is both useful and non-intrusive. What is also very interesting about this data, is that one of the only semantic web pieces of software to be used is Avogadro, which shows a significantly lower usage here than in the molecular modelling software. This is probably because even though the work produced by (Hanwell et al., 2012) makes it clear that it is a semantic platform, the software website doesn't market it as a 'semantic' product, but rather as a molecular modelling package. This suggests that perhaps quite a lot more people use it but wouldn't categorise it as 'semantic'. In total 46 participants said that they had used Avogadro, with 16 saying they used it often and 30 saying that they used it occasionally. This is a prime example of a product that uses semantic web technologies non-intrusively; as it uses the Chemical Markup Language (CML) and uses that to facilitate interoperability between other software packages and to add semantic meaning to the data. Given this piece of software has proved popular in two categories it is definitely one to be considered for the ELN environment.

4.4.1.8 Organic Synthesis, Nanostructure Modelling & Chemical Kinetics & Process Software

These categories not only showed very little usage overall, but the software packages listed within them also showed a very low usage across the board, suggesting that certainly in this sample of chemist this is not a branch of software that is widely used. Due to this, these graphs have not been included; however the full results can still be examined in the raw data which is detailed in Appendix B.

4.4.1.9 What Piece of Software Would you Most Like to be Created?

This question was asked at the end of the survey, and only about a quarter of the participants answered. However, the responses given were very interesting. There were some calls for some very specific pieces of software such as taking .xyz files (atomic position and species) to figure out functional groups, connectivity and atom classes.

Several participants requested better drawing and visualisation tools, including a desire for a tool that facilitates all types of chemistry drawing. Other requests lead towards

[10] `http://www.chemspider.com/`

[11] `http://webbook.nist.gov/chemistry/`

[12] `http://www.nist.gov/pml/data/`

the desire for more integration. The vast range of software packages that were found when creating this survey illustrates how many different software offerings there are for chemists, and based on the results and high levels of usage shown for different types of software in different categories, some chemists must use multiple types of software. This would lead to files and images of different formats, and several participants said that they would like tools to be better integrated, with tools to convert between file formats, and to produce rich formatted output files from different types of input files. One participant answered that they wouldn't like any more tools to be created but that they would like the pre-existing tools to work together better. Another said that they would like an integrated set of tools where its easy to search databases and perform calculations. These results suggest that there is a calling for tools that perform multiple (ideally all) chemistry functions rather than individual packages for each different type of functionality.

This suggests that there could be a need for a tool that looks to integrate different chemistry tools and where possible link with existing tools. The results of this survey make it clear that chemists do use a lot of software packages, and many chemists identify themselves as working in multiple areas of chemistry, which also mandates using multiple different pieces of software. Additionally, the results of the literature search conducted in Chapter 2 found evidence that appropriate domain software was a feature to include in an ELN, thus this survey has allowed us to identify which popular open source pieces of software could be used to create domain based add-ons in an ELN.

4.4.2 Focus Group Results

The focus groups were conducted second and had a total of 24 participants take part in them, 6 physicists, 4 biologists and 14 chemists. There were some limitations in availability of participants and timing, and the overall makeup of the 24 focus group participants are listed below; links to the full set of questions, ethics applications and verbatim transcriptions are all listed in Appendix C. The results have been broken down into the different areas of questions that were asked in the focus group, and these categories will be used throughout the thesis in Tables 4.2 and 4.3

- Chemistry Focus Group 1: 8 Participants: G, H, I, J, K, L, M &N
- Chemistry Focus Group 2: 6 Participants: S, T, U, V, W & X
- Physics Focus Group: 6 Participants: A, B, C, D, E, & F
- Biology Focus Group: 4 Participants: O, P, Q & R

4.4.2.1 Recording Notes

The focus group began with a general discussion about how the participants recorded their notes, and what types of information they recorded in different circumstances. There was a resounding agreement that all participants used their paper lab notebooks for recording their notes, although there were some differences aside from that. The biologists seemed to rarely employ any software for taking their notes. Most of the notes were made by hand and standard word processing and spreadsheet software such as Word and Excel were employed for any writing up, calculations and graph production. Whereas the chemists and physicists both mentioned using a computer alongside their paper lab notebooks, the physicists showed a trend of using their lab notebooks for notes, and their computers for data files, with the exception of Participant F who used a computer for everything. The chemists also followed similar trends to each other, but with varying level of organisation. The computational chemists tended to use some software, with sporadic use of their lab book and random scraps of paper. Whereas the 'wet' chemists had stringently organised different lab books for different things, with the additional slight anomaly of Participant J who liked to use blogs and Word documents alongside the paper notebooks.

The types of notes taken and the manner in which they were taken differed depending on the situation. For experiments inside and outside the lab there was a wide use of the lab book, generally to take down observations and write down initial values. The chemists used a few different software packages, such as XMGrace, and also the standard Excel and Word. The sheer amount of different data that they recorded ranged from energy values, to temperatures, masses, observations, schemas and protocols. This showed very similar results to Reimer and Douglas (2004) who conducted their studies 12 years ago, illustrating how the information recorded in the scientific record keeping process remains much the same. These results have made it very clear that the different types of chemists have different needs for recording their notes, therefore meaning that any ELN would need to cater to the different types of chemist as opposed to just trying to make generic features available. Thinking about their work tended to lead to writing notes in a notebook / lab book or often on scraps of paper which sometimes proved difficult to organise or locate at a later date. Despite the contrasting ways of organising their notes, each scientist had their own method of organisation that they set up and stuck to. Shankar (2007) concluded from his studies that several participants had set up their own 'templates' or attempted to standardise their data entry in certain ways. The scientists involved in these focus groups seem to be doing the same thing, suggesting that offering them a way to do this via an ELN environment could be very popular.

The participant's answers illustrated that some tasks definitely had a paper preference, such as looking at literature. Most participants said that they printed out their research papers and highlighted them and made hand written notes. Some typed up notes or made them in their reference managers associated with the appropriate papers, but even they said that for the longer papers or ones with the more complex theory they would print them out. There was a general agreement that the participants found it hard to read papers on the screen and preferred to have a physical copy. Whereas whilst the physicists liked to write out their equations on paper to solve them, the calculations side of things tended to lend itself to software and computers. Most participants in both chemistry and biology said that they used Excel, and other software packages such as XMGrace, Spartan and Wolfram Alpha were mentioned. Generally, these packages were used to either check the equations or create graphs and figures out of the raw data.

The area where the participants mentioned using software the most writing up their work, either for publications or to create presentations or posters. There seemed to be little notion of writing up experiments or procedures neatly just to make a neater copy of the lab book (although a few participants did this), but they all used software to write up reports, conference papers, publications and any weekly presentations or reports that their supervisors required them to produce. These were written up in Word, LaTeX or Lyx (a graphical editor for LaTeX), or using presentation software such as PowerPoint, with graphs and figures produced from the software packages mentioned above, although Excel seemed to be the most popular.

4.4.2.2 Organising Notes

The participants were then asked how they organised their notes. This proved to be a personal endeavour, fitting with Shankar (2007)'s conclusions that taking notes and creating a standard method of data entry was personal to the different scientists, as it would logically follow that then organising those notes would be equally personal. Most participants record data such as the date and title, but some of them organised their work chronologically and others categorise it in different ways. Generally, the lab book is chronologically organised, although they might have different books per category. Some participants categorise their data by experiment type, or subject area if they are doing interdisciplinary work or even work that spans across several sub disciplines of their science, or per project if they are working on several at once.

The chemists, in particular, also had different knowledge organisation structures in place for linking between notes. The biologists deliberately dated all their work such that different documents and lab book entries created on the same day could be located later when looking through their work. The physicists liked to organise their work by categorising pages in their lab books and electronic documents under an experiment or project number so that they could find the associated materials for one experiment or project at a later date. The chemists seem to be the only scientists who use codes to link between their lab books and their computer based notes. They have almost all created custom codes to assign to electronic documents such that they include information like a lab book page number that is associated to the electronic data file, to aid in locating the appropriate pages in their lab book when looking at their electronic work. Some of the chemists also mentioned using the contents page of the lab book to aid with searching later on.

This illustrates that the scientists do all try and organise their notes to varying degrees to try and make it easier to locate specific pieces of work at a later date, suggesting that they must look back at their lab book after creating the initial entries. This correlates with a conclusion from Shankar's study where a participant stated that they liked to write up everything into his lab book so that he could go back to it at a later date. This also suggests that even if not all scientists see the point in digitising all their work, they clearly re-use their paper work, and also have a requirement to organise it in such a way that they can find things easily.

4.4.2.3 Use of Technology in Note taking & Lab Equipment

The group then moved on to discuss how other pieces of technology were used to aid in their note taking. This area made it very clear that we have transitioned into a modern age of technology where it is common to use a smartphone as part of the working environment. Several participants said that they used their smartphones to take photographs of their work, their experiments, results, slides at conferences, discussions on whiteboards, or even full on recording of an experiment procedure. It also seemed commonplace to email oneself with reminders or links, and there was a use of smartphone calendars to organise meeting plans rather than a paper based diary. Additionally, there was also the usage of cloud storage and note booking software such as Dropbox and Google Drive, although there was still an obvious element of more traditional ways of transferring data such as USB sticks, email or memory cards.

When questioned as to what equipment they could take into the lab this differed across the different disciplines. The biologists didn't seem to have any restrictions, although one biochemist mentioned that they had to be careful about bringing in equipment from the outside in case of contamination. The physicists who frequented the cleanroom weren't allowed to take in bits of equipment in case they contaminated the environment, whereas the chemists didn't really have restrictions but self-imposed them on themselves, as they didn't want to contaminate their technology with chemicals.

4.4.2.4 Data Storage, Backup and Linkage

Following on from the previous theme of needing to link together the different resources, the participants were asked where they stored their data, how it was backed up, and how they linked together their paper and electronic notes, if at all. The main way that the biologists linked their notes together was by dating everything so that things could be joined back together later. The physicists also linked their data together by using dates, but also by categorising work under experiment numbers. The chemists seemed to take a slightly more methodical approach by using codes to link together lab notebook entries and electronic data files. However even this was done in a different way by almost every chemist participant, the chemists noted using lab book pages, initials and sample numbers, in addition to physically noting down file locations in their paper lab book to show themselves at a later date where the associated electronic data files were. This furthers the theme of note organising being very personal. It also shows that there is a clear need for linking between paper and electronic notes, and that this needs to be taken into account for any platform created. It also suggests that there is a willingness to both use technology and create links with electronic records, provided they are seen as useful to have.

Several participants also showed a high amount of concern for backing up their electronic notes. Most of them had their electronic data stored in various places, on their own computers, on university computers, on shared drives, and cloud storage. In Southampton University there are shared drives for each department and the students have clearly taken advantage of that, and have set up group folders on these drives. However, despite setting such a high importance against backing up their electronic work, most of the participants showed significantly less concern about backing up their paper work. All the physicists and most of the chemists and biologists said that they didn't backup their paper lab books at all. A few chemists have the lab books with the replica pages you can tear out, but even those admitted to either not having torn out their latest lab books, or having left the backup pages in the lab which somewhat defeats the point of having them as backups. One of the biologists was made to photocopy their lab pages and give them all to their supervisor, however this was clearly at the supervisors' request.

This lack of concern doesn't correlate with the effort that some of the participants put into organising their paper lab notebooks, or linking together their paper and electronic work. The organisational efforts that they make to ensure that they can find things in their lab books at a later date, and link back to their lab book when looking at electronic files suggests that the material of the paper lab book is re-visited. However, the lack of effort to back up the paper-based work suggests that the scientists wouldn't be too concerned if they couldn't go back to it later. It could be that the participants see more of a risk to their electronic work, as there are many ways in which electronic files can be corrupted or deleted, but given the evidence suggesting that scientists do require going back to their paper lab notebooks, further steps need to be taken to improve the digitisation of the scientific record, such that these important notes are not lost due to lack of paper backups.

4.4.2.5 Intellectual Property

An area that needs to be broached when considering the concept of a cloud based notebook is that of Intellectual Property (IP), as once data is stored outside an individuals' private machine there is a higher potential for it to be viewed by others. There was a high level of contrast in the discussions about IP both between the participants from different disciplines, and within the separate discipline groups. The main difference between the chemists and physicists seemed to be that those with industry sponsors were more concerned about IP than those without. The industry-sponsored students were required to get sign off on their presentations and send in reports, and in some cases weren't permitted to use cloud software because the data needed to be kept secure. The students who were funded through other means such as research councils seemed to use cloud storage quite happily and didn't require any sign off. The biologists were different in that even though they weren't working with industry, they were in some cases working with animals, which are considered sensitive, and that associated data was stored on a hard drive in a locked drawer. They told horror stories about having to stand by their posters at conferences to make sure that no one could take a picture of their work, and how they weren't allowed to present their work until it was at a certain stage. A few of the chemists equally had stories about how members of their group had been 'scooped' and that their supervisors had heightened their security measures after that.

Potentially some cloud based ELN environments won't suit every need, due to concerns about data privacy; for example, some of the participants with industry sponsors weren't allowed to use Google Drive or Dropbox; with the physicists noting that this was because Google Drive did not conform to the European Environment Agency Privacy Policy regarding personal data [13]. However, there seem to be enough scientists who aren't concerned about IP and who already have a substantial amount of their data out there in the cloud that it should still have a viable audience, and if it were to prove successful, work could be done to produce standalone versions with the same functionality that operated within a university network for example.

4.4.2.6 Collaboration and Sharing

The notion of collaboration and sharing seems to be a popular one among the participants. A lot of the participants collaborate either with people, groups or companies outside the University, and within their own research groups at the University. Therefore a collaborative piece of software would surely aid with this. Despite the concerns about IP and sharing their data outside the university, it seems like there is a very open nature towards sharing with group members, with participants looking at each other's lab books

[13] https://www.eea.europa.eu/legal/privacy/privacy-en

and some supervisors keeping records of lots of students old lab books for future reference. Having a collaborative ELNs would enable this sharing to keep going, whilst making it easier and more accessible in the long term; and this sharing, particularly for feedback was dubbed very useful by the participants.

4.4.2.7 Reference Management

Participants were asked about their reference management procedures, and they all said that they used a type of reference management software. The popular ones mentioned were EndNote, Papers, Mendeley, Zotero and BibTeX. Reference management is clearly area that all the participants are prepared to use software for, and therefore would definitely need to be integrated into any ELN environment that was created.

4.4.2.8 Scenarios

The participants were presented with three scenarios to discuss:

- *Imagine you're trying to locate a piece of work or some notes from 6 months ago, how would you locate these notes, and how would you locate your data?*
- *Imagine there's a fire in your lab and all of your paper notebooks are destroyed, how much of your work would be lost and how could you go about recovering this work?*
- *If you fell under a bus tomorrow, and were indisposed for a while, how would your supervisor / industry sponsor / colleagues access your work*

These scenarios were met with differing reactions. Some participants seemed totally unconcerned at the prospect of losing all of their lab books, and didn't seem to think that it would take too long to reproduce what was needed, as a lot of it was information that was only useful in the moment, or a list of things that didn't work and were improved upon. Whereas other participants greeted the suggestion with cries of 'I'd be ruined', 'a nightmare', 'I might as well stop my PhD now'. Particularly with reference to the idea of their labs catching on fire several participants seemed more concerned at the idea of losing their lab samples or compounds. Suggesting, that perhaps their lab books would not be the thing they would be dashing back into the flames to retrieve. However, this doesn't entirely correlate with the suggested evidence detailed in Section 4.4.2.4, that scientists do go back to their lab books. Otherwise, why would they put so much effort into organising them, and in the case of the chemists create linking codes between their paper and electronic documents.

With regards to searching through their data, most of the biologists and physicists said that they would flip back through their lab book by date to locate work from previous months, and would likewise search by date through their computers to find the appropriate data files or would search by filename. The chemists gave a wider variety of approaches, with some participants saying that they would also search back chronologically through their work, and others saying that they would search using their custom codes to find electronic and paper notes. Some chemists said that they would use their codes to find associated lab pages and electronic files so that they could retrieve the necessary notes and data for a specific experiment. Additionally, there seems to be a system of sharing within research group's setup for most of these participants, but less so directly between some supervisors and PhD students, or even certain industry sponsors. A majority of the participants (aside from a few with heavily involved supervisors) admitted that their supervisors would really struggle to search through their work to find something specific or even to access it in the first place. This continues the earlier theme about showing less concern towards backing up paper-based work or having any contingency plans set aside for them. Furthermore, it shows that some of the participants are not considering maintaining their work for other researchers; which is concerning with regards to maintaining the scientific record for prosperity and future work.

Seemingly despite being aware of the potential issues that scenarios like this could cause, the participants do not consider them likely or serious enough to merit pre-emptive preparation, apart from circumstances where their supervisors have put procedures in place. This links to the social influence factor of UTAUT; whilst organising their work isn't the same as adopting a new piece of technology, it illustrates that significant role that social influence can play in how a person behaves and what practices they put into place with respect to their work. This suggests that more than just a software based solution is required to improve the management and digitisation of the scientific record. Currently, despite many scientists undertaking collaborative endeavours, even for digitised work there is a lack of preparation for ensuring that their work can be accessed by others at a later date. Some supervisors have put better practices into place such as requesting copies of students work, and mandating certain backup procedures; but in places where this social influence has not occurred, desirable procedures have not been put into place. More work needs to be done to improve management practices alongside further digitisation to ensure that the scientific record is well preserved for the future.

4.4.2.9 Current usage of Paper and Electronic Devices

The feedback from the focus groups indicated that researchers work in very different ways, with different working patterns and that the way in which they create their notes is highly individual. There is a mix of paper and electronic usage across the different disciplines, Table 4.2 details a summary of this based on the results described in the previous sections.

Tasks	Biologists		Chemists		Physicists	
	Paper	**Electronic**	**Paper**	**Electronic**	**Paper**	**Electronic**
Planning Experiments	Lab Book	Not Used	Lab Book	Not Used	Lab Book	Not Used
Recording Experiments	Lab Book	Not Used	Lab Book	Data	Lab Book	Data
Reviewing Literature	Print papers / write notes	Reference Manager	Print papers / write notes	Reference Manager	Print papers / write notes	Reference Manager
Reflecting on Notes	Lab Book	Not Used	Lab Book	Google Tasks	Lab Book	Google Keep
Organising / Linking Notes	Lab Book by Date / Contents Page	Paper & Electronic Notes linked by Date	Lab Book by Date / Contents Page	By codes (linking to Lab Book) & by sample / experiment	Lab Book by Date / Contents Page	By codes (linking to Lab Book) & by category / experiment
Searching Notes	Flip back & Search by date	Not Used	Flip back & Search by date	Sort by date / search by codes or keywords on computer	Flip back & Search by date	Sort by date
Processing Data	Solve Calculations	GraphPad / Excel	Solve Calculations	Wolfram Alpha	Solve Calculations	Excel / XMGrace / R / Spartan / PyPlots / CSV
Writing Reports	Not Used	Word / Powerpoint	Not Used	Word / LaTeX	Not Used	Word / LaTeX
Technology in the Lab	N/A	Phone pictures / recordings	N/A	Phone / Camera pictures, Emails, Blogs, USB Sticks	N/A	Phone pictures / calendar, Emails
Archiving & Backup	Mostly no backup (some photocopies)	University Computers / Shared Drives / The Cloud / Hard Drives	Mostly no backup (some use carbon pages)	University Computers / Shared Drives / The Cloud / Hard Drives	No backup	University Computers / Shared Drives / The Cloud / Hard Drives
Intellectual Property	N/A	Secure data like animal data kept on hard drive in locked draw	N/A	No cloud software for industry sponsored students	N/A	No cloud software for industry sponsored students
Collaboration	Lab Book	Not Used	Lab Book	Shared Drive / Group Folders	Lab Book	Shared Drive / Group Folders

Table 4.2: Focus Group Results of how scientists use paper and electronic devices in lab process, adapted from (Kanza et al., 2017)

Table 4.2 shows that for the biologists in particular there are several areas of work (e.g Recording experiments) that are conducted purely as paper based tasks, and whilst the chemists and physicists do not have any purely paper based areas of work, they still use paper alongside electronic devices in almost every area. This, combined with comments about how much easier some participants find paper to use suggests that an ELN would need to both fulfil the Effort Expectancy and Performance Expectancy criteria of UTAUT to convince scientists that it would be both easy to use and would improve their performance to persuade them to change their current practices.

Furthermore, both the biologists, and to a lesser extent the physicists, showed uniformity in their approaches to their work. The chemists however showed a higher level of diversity by using a wider variety of technology to support their work, and by creating different organisational structures with which to search and categorise their work. It is also worth noting that the only work based task displayed in Table 4.2 that is purely electronic is writing reports. The participants have demonstrated that when it comes to writing up their work for publication or presentation that software is always used. Therefore, an ELN would need to cater to scientists' report needs, as well as facilitating their note taking; which further reinforces using some of the existing features of Electronic Notebooks in an ELN, or building an ELN on top of existing notebooking software.

4.4.2.10 ELN Opinions & Experiences

A few participants had used ELNs before such as LocalWiki, LabTrove, Blog3, BioBook, Enovalys and one in industry that they couldn't remember the name of. None of them seemed to have committed to these in the long term. The experience of the industrial ELN was by all accounts not a pleasant one, and whilst the other ones seemed to have provoked less distress, it seemed like they had only seemed useful for certain endeavours. Participant S used Enovalys, and found that it was very useful for their inorganic work, but didn't provide the required functionality for their transport runs. Equally, the participants who tried LabTrove and Blog3 found some of the elements useful but it seemed to depend on the situation, and they all defaulted back to Word documents in the end. One participant stated: 'People don't use them because they don't really integrate into the scientific workflow in a way that makes them feel like it's making their life easier' which fits into the Effort Expectancy (how easy it is to use) factor of the UTAUT. Clearly the scientists need to feel like an ELN would be easy to use before they would be willing to consider making that step. Additionally, some strong reactions to the idea of ELNs suggests that the addition of an ELN to the network would indeed cause disruption as denoted by ANT.

There seemed to be contrasting opinions about what an ELN could do for you, some participants seemed to think that it would make their life easier and would be a good idea if it had the required functionality, whereas others took the opinion that there wouldn't be much point in using one and that they would still prefer to use paper. However, some very useful suggestions were given when they were asked what they would want from an ELN; these suggestions combined with the needs elicited from the rest of the discussions have been broken down into the following categories:

Organised

Participants want their ELNs to be organised, indexable and searchable. They want to be able to tag and classify their notes so that they can be suitably categorised, and to be able to store metadata. The discussions with the participants about how they organise their notes have shown that scientists like to categorise their notes, and particularly amongst the chemists, like to be able to link between pieces of work. Additionally, users also requested the ability upload and link files and images to their notes as well as their data files. Tagging notes, storing metadata, and linking between documents are pieces of functionality that can be achieved by using semantic web technologies; illustrating that the 'Rich Semantics' feature of an ELN proposed by the outcomes of the literature search in Chapter 2 is indeed a useful feature to include.

Simple & Personable

Users want their notes to be personable, so that they can create them and organise them the way they want. The participants have demonstrated how personal an endeavour writing their notes is. An ELN would need to allow users to organise and create their notes how they want, with the ability to create folders and structures and assign different categorisations if desired. Additionally some participants mentioned a reluctance to create their notes electronically initially because it felt too formal, and said that they liked to customise their notes with different colours and layouts. Participants also wanted an ELN to maintain the affordances of paper, saying it should be as easy to write in as a paper notebook. This links back to the Effort Expectancy criteria of UTAUT where in order to adopt an ELN users would need it to be easy to use.

Flexible & Templated

The ELN should facilitate many different types of experiments, with a wide range of templates to aid in data entry and organisation, but that still offer customisable layouts, and do not force all experiments to conform to the same templates.

Domain Based Functionality

One of the outcomes in Chapter 2 was that an ELN would need to contain the appropriate domain based functionality, and requests were made by the participants that matched with this conclusion. Participants requested the ability to perform calculations and handle formulas and equations, and to handle scientific diagrams. The results of the survey alongside these focus groups illustrated that scientists use software to model and analyse their data, as well as using different software programs to interact with different machines. An ELN would need to either contain these domain based features or interface with the programs that scientists regularly use such as ChemDraw.

Collaborative

The participants also stated that they would want an ELN to be collaborative, and the discussions demonstrated that a majority of them took part in collaborative projects and or collaborated with others as part of their work. This feature was also proposed as necessary to incorporate into an ELN in Chapter 2.

Cloud Based

Despite mixed opinions about cloud systems with regards to data protection and security, some participants professed a desire for an ELN to have Dropbox like features.

Electronic Notebook Based Functionality

Finally, some features that already exist in note booking software were requested, such as 'diagram creation and features like Word and Excel', and the ability to link to reference managers. This is unsurprising as a majority of the participants already use Word and Excel to create their work and want their features to exist in an ELN. This further illustrates the overlap between what users would want from an ELN and what they want, and indeed already get from an Electronic Notebook. These features will be further considered with respect to the Dial-a-Molecule results to see how closely they correlate.

4.4.3 Dial-a-Molecule Surveys

This thesis involved a collaboration with the Dial-a-Molecule group to access their ELN survey and study data; this was analysed and written up alongside the studies from this project in (Kanza et al., 2017) to form a proposal for the prototype ELN. The surveys they conducted identified a number of features that users said they wanted from ELNs. These have been broken down into how they were prioritised by the users in Figure 4.8.

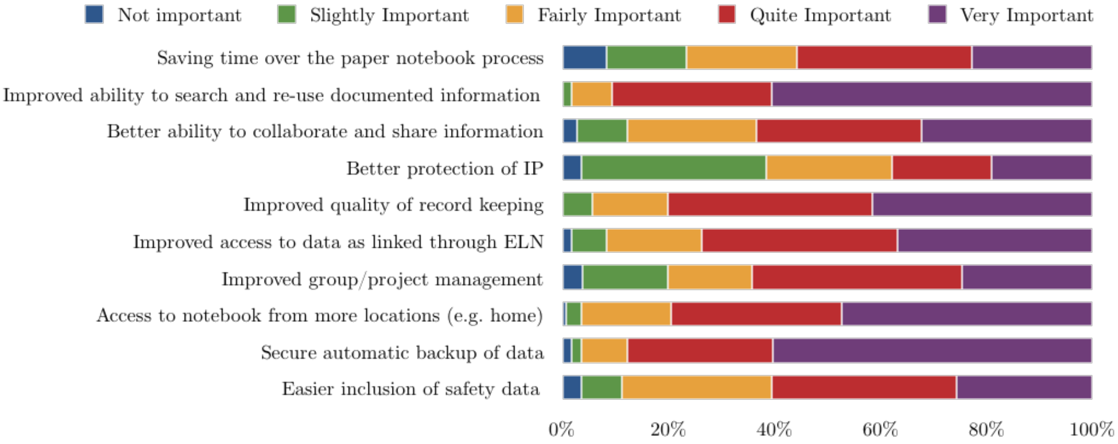


FIGURE 4.8: User priorities elicited from the Dial a Molecule iLabber Pilot Project: Potential Uses of ELNs in Academia Survey from September 2011, adapted from (Kanza et al., 2017)

These priorities are also detailed in Table 4.3 to illustrate which priorities map to which user desired features from the other Dial-a-Molecule survey and the other user studies detailed in this chapter. These priorities illustrated that secure automatic backup of data and the improved ability to search and re-use documented information were considered as the most important. This fits with the user behaviour demonstrated in the focus groups, in that the participants were very focused on backing up their electronic work, although currently most of them didn't really backup their paper based work. Additionally, the participants in the focus groups and lab observations demonstrated how some of them had organised their notes to try and make it easier to find certain pieces of information, but equally demonstrated the vast amount of data and pieces of work that they needed to coordinate; highlighting the need for better searching functionality. Furthermore, this fits with the Performance Expectancy criteria of UTAUT in that the features identified as the highest priorities are areas that can be vastly improved using technology compared to using paper. This suggests that the users prioritise features that would improve their current abilities rather than just replacing what they do with an electronic solution.

The features that the users wanted that were identified in the Dial-a-Molecule survey alongside the features identified by the initial user studies in this project have been categorised in Table 4.3. The Category column details the different areas of note taking and work identified by the focus groups that were explained in Table 4.2. The Desired features column details the features requested by the user in the Dial-a-Molecule survey which have been grouped according to which note taking category they fit into. The Priorities (DaM) / Addressing Barriers column links to both the main adoption barrier categories detailed in Table 2.1. These show which priorities are addressed by each set of features, and which barriers could be mitigated by them. These results form the body of the requirements for the proof of concept ELN which is fully detailed in Chapter 5.

Category	Desired Features	Priorities (DaM) / Addressing Barriers
Recording Notes	Simple to install Personalisable Post-it notes TODO lists Create default values Easy to write in as a paper notebook Facilitate different experiments Range of experiment templates	55.6% - Saving time over the paper notebook process is important Barrier: Ease of use
Organising Notes	Indexable / Highlightable Contents Table/ Overview screen / Timeline Spellchecker Tag / classify notes and experiments Store metadata Use of standard vocabularies (ontologies/measurement techniques)	80.2% - Improved quality of record keeping is important

Searching	Keyword / Filtered Search Data traceability Advanced searches by Chemical Structure Include reactions schemes in search results Voice Searches Sortable Results	90.6% - Improved ability to search and re-use documented information is important
Linking Data	Upload / link files, images and data files to notes Link between different notebooks Link to reference managers Dropbox-esque features (automatic data update) Automatically link to external chemistry resources	73.6% - Improve access to data as linked data through ELN is important Barriers: Data Compatibility & Portability
Writing Reports	'Generate Report' button to generate a publication ready report Integrate and store different types of documents (Excel, Word, PDF, Pictures, Handwritten notes) Copy sketches into notebook Paper notebooks integration Digital pen integration Migration tools Export functionality	Barrier: Software & System Integration and Compatibility
Performing Calculations & Scientific Functionality	Perform calculations, formulas and equations as easily as paper Create sketches and diagrams Recognise a chemical when entered Risk Assessment Templates / view electronically Flags for dangerous chemicals Index of COSHH materials Global database of chemical values Notifications for approvals Sign off entries to make them non editable	60.4% - Easy inclusion of safety data is important
Use of Technology in the Lab (Accessibility)	Web Based / Platform Independent Tablet / Smartphone Compliant Text recognition, drawing and photo capabilities Usable in the lab like a paper notebook Voice capture Built- in language for extensibility	79.3% - Access to notebook from more locations is important Barrier: Access
Archiving & Backup	Secure storage, backup and archives Downloads / Printing	87.8% - Secure automatic backup of data is important
Intellectual Property	Secure access Different access levels for users	37.8% - Better protection of IP is important

Collaboration	Shared files / notebooks Standard list of instruments and reagents Link related people and notebooks Coordination for Open Source and Access Sign up and 'get involved' pages Configurable stand-alone to act as portals for projects and landing pages for collaborators Enable users to find out who is working on similar molecules of reactions (requires inbuilt understanding of molecules)	63.2% - Better ability to collaborate and share information is important
Project Activities	Recent Activity feed with notifications Page statistics Bulletin Boards Moderate comments	64.1% - Improved Group / Project management is important

Table 4.3: Desired user features from collated user study data, linked to related priorities and barriers from the Dial-a-Molecule surveys, adapted from (Kanza et al., 2017)

These combined results highlight the need for the features identified in Chapter 2, alongside many more different pieces of functionality. It is clear that there are software features that are desired by scientists to improve their current software offerings even if they do not all currently wish to use ELNs. Furthermore, the desired features link with the adoption barriers presented by both literature and Dial-a-Molecule and BioSistemika surveys. The priorities elicited from the surveys mostly illustrate specific improvements that the users would require in order for them to consider using ELNs. This strongly links with the Performance Expectancy element of UTAUT as scientists would need to feel that an improvement to their current offerings was being made in order to adopt an ELN.

4.4.4 Participant Observation Results

The participant observations were conducted third, and four different lab observations took place; the following labs that were observed, and the participants involved in each lab observation are listed below, and links to the full observation notes and ethics applications are all listed in Appendix D.

- Crystallography Lab - Participants W, AF & AG
- Molecular Chemistry Lab - Participants S, T & AH
- Organic Chemistry Lab - Participants AI, Y & Z
- Inorganic Chemistry Lab - Participants AA, AB, AC, AD & AE

4.4.4.1 Crystallographers

This lab differed from the conventional chemistry lab setup; there were two main rooms, the first room was the prep room which the scientists used to prepare their samples and make initial notes before progressing into the main crystallography labs. The main labs consisted of a few big machines in it hooked up to a few computers, with no visible lab coats or test tubes. Figure 4.9 shows the prep room and main lab room.

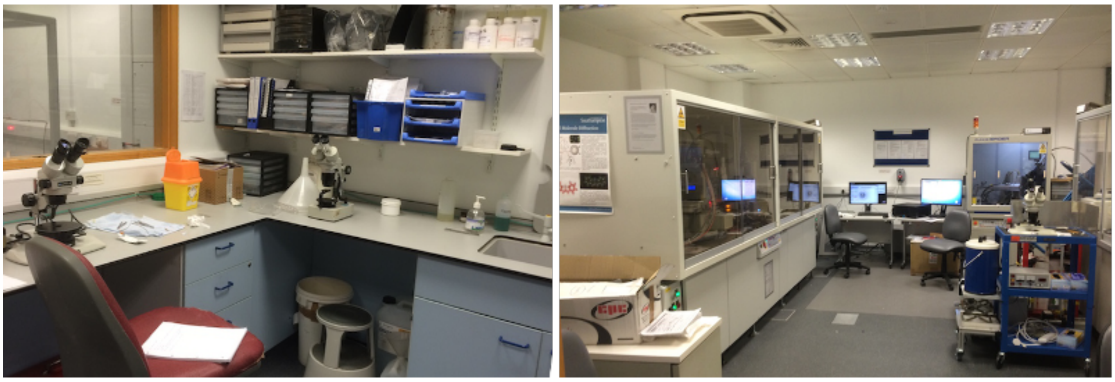


FIGURE 4.9: The Prep Room (left) and Main Lab Room (Right) in the Crystallography Labs

Participant W from the chemistry focus group was part of this research group and the observations looked at the participants doing their regular crystallography work. This work begun in the prep room, where they prepared their crystal sample under the microscope. Participant W took notes in their sample book of information that will need to be input into the computer in the main laboratory. The information associated with the crystal sample that has been sent over by the crystallography agency that this group work with is paper based. The forms have been filled in by hand and all the information is on sheets of paper with the test tubes of the sample crystals.

This immediately shows that there is a high element of paper-based work involved in these experiments, including the information prepared and sent over by the crystallography agency. This could suggest that with larger establishments still using paper, that these paper-based processes remain normalised. Furthermore, receiving information on paper rather than being able to look it up somewhere electronically means that at some point manual data entry will have to occur. In this instance, it occurs twice, as some information is copied from the paper based sample sheets, into a paper sample book, and then input into the computer in the main lab.

When asked if Participant W would consider making these notes on a laptop, the response was: *'it would be a bit awkward to have a computer in this room and a laptop in the prep room, I use my book for a quick reference'*. This illustrates both the accessibility barrier in that Participant W felt it would be difficult to use additional technology in the lab, and links to the Effort Expectancy criteria of UTAUT, as in this setup a paper lab notebook is deemed easier to use than an electronic device. Once the sample had been isolated to Participant W's satisfaction, they progressed into the main lab.

Participant W opened up CrystalClear software, created a new project and entered in the sample information. The sample was then placed in the diffractometer and some more information was written down in the sample book; again illustrating how paper is used for quick note taking. The software has an on-screen ruler so that the crystal can be measured. The molecular formula from the sample sheet is then entered into the software; which produced some preliminary information, which was also written down in the sample book.

Participant W then needed to fill out another paper based diary, which the lab occupants all use to record the experiments that have been run on what day on what machine. Again, these paper-based processes seem to have been normalised for this lab, even though if this had been filled out electronically it could potentially be accessed from locations other than directly in the lab, and it would be easier to backup. Whilst Participant W was waiting for the machine to be finished, the software crashed, which they said was a common occurrence. This demonstrates that sometimes software can be very unreliable, and suggests that perhaps some of the paper-based processes could partially be due to a lack of faith in the software. Facilitating Conditions are also an important part of technology adoption as denoted by UTAUT, and if users do not feel that their technology is well supported with a suitable technological and organisational infrastructure then they will be unlikely to consider adopting it.

Due to the software crash, Participant W had to go and speak to Participant AF who was remote desk-toped into the machine; and together they were able to restart the system from Participant AF's computer. Following this, Participant W took some more notes in their sample book and said that the crystal can now be left for the next 3 hours. When Participant W returned to their office, the rest of the researchers in there were all sat at their computer desks doing different things. Participant AF was creating some 3D structures, and Participant AG was looking at emails. Participant W logged into a crystal portal and entered in the appropriate information from their paper sample book, which will then have the data files zipped together and be added onto the record later. Participant W then started another piece of software called CrysAlisPro, which is used for processing. Unfortunately, Participant W explained that it doesn't quite transfer all the data from CrystalClear, which is why Participant W needed to note down some of the figures in their sample book. Then another piece of software called Olex^2 was opened, which is for visualising the data, to provide the molecular structure and detail how 'good' the data collected is.

These practices show a very disjointed set of software packages that do not work together very well, and require additional work to be done to transfer data and information between the different packages, explaining to an extent why Participant W used a paper sample book to quickly note down these values. When discussing this with Participant W, they mentioned that they hope to start using a unified system soon that will bring together all these different pieces of software into one system.

Discussing Participant W's personal work aside from the crystal sample work lead to a discussion about how a project has developed through their group in an excel spreadsheet that contains tables of other crystal data and grids of substitution:

Observer: 'Why wouldn't you just use Google Sheets and use the computer in the lab'?
Participant W: 'They are attached to a diffractometer so I wouldn't want to use them'.
Observer: 'If you had another computer in the lab with Internet access and not attached to a machine would you consider using Google Sheets and use it within the lab?'
Participant W: 'Yes that would be useful, especially as it has been passed down from different people who all kept their records in different ways'.
Participant AF 'These projects started and they didn't know it would go further'.
Observer: 'Do you think it would be a good idea to have a separate computer in the lab with internet access?'.
Participant AF: 'Why wouldn't you just use the computers in the lab?'
Participant W: 'Using a browser wouldn't be an issue'.

This suggests that this area that has not been thought about much, and Participant W described a laborious process about how they sit down side by side with their colleague and manually copy each other's copies of the spreadsheet. This suggests that perhaps these scientists aren't so much against the use of technology in their work, but that they haven't necessarily thought of alternative methods to the ones they currently use. These discussions have elicited a lot of disjoint activity in this area, both between needing to use three different software packages for one simple sample, and the need for manually writing down and entering data into a paper sample book to transfer between software packages. This suggests that unified software packages combined with some insight on the broad range of solutions available could prove very useful to lab environments such as these. These observations link back to the final question of the survey regarding what software chemists would most like to be created, as several survey participants said that they would like tools to be better integrated, or that they would like existing tools to work together better rather than new ones being created. This leads back to using domain based add-ons on top of an existing note booking piece of software, to incorporate the software functionality that scientists already use, combined with integrating the domain based tools they need such that more work can be done using one system.

4.4.4.2 Molecular Chemists

The molecular chemists also had multiple lab rooms in which to work, although unlike the crystallographers these rooms weren't joined together for preparation and experiment.

The three rooms that were looked at during these observations were located on different floors of the same building and housed different equipment. The first room was the main lab, which was set up for conventional bench chemistry, with workbenches for the scientists to weigh out their materials and write in their lab books, and fume cupboards to conduct the experiments in. There were many hazardous chemicals present, and white coats and safety goggles were mandatory equipment for entering the lab. This lab also had some desktop PC's connected to some machines. Directly adjoined to this room was the main computer room where the students had their desks, and the third room was the NMR room which held specialised equipment.

When observing the main lab room, one of the participants in this lab, Participant AG said that they had used LabTrove for some crystallography experiments and quite liked it. Interestingly after comments in the focus groups about not wanting to take laptops into the lab, this lab seemed to have its own personal laptop. Participant AG had their phone in the lab and used both their lab notebook and the lab computer. They put their samples into the fluorometer (which measures fluorescence decay of a compound); this machine was hooked up to a desktop computer, and this computer was only used to process the results off that specific machine. Participant AG used a USB drive to transfer the results to their lab computer, and when they were queried as to why they were using this method of transferal rather than using any cloud based software, they explained that the computer in question didn't actually have internet access. Further enquiry suggested that there was a concern that the computer would be more accessible to virus's if it had internet access.

Participant T (from the focus group) was then observed in the NMR room which was another room consisting of large machines and computers hooked up to them. I was warned that the computers and indeed any pieces of technology needed to be kept a certain distance from the machines as they contained powerful magnets. The computers ran iconNMR and ACD labs, which were used in conjunction with the lab book for these experiments.

After returning to the computer room, a discussion was had with Participant S (from the focus group also) who was working at their computer because their experiment hadn't worked. They explained they were working on an abstract for a conference, which was being written in word, and gave a run through of the notes and files on their computer and in their paper lab book. This illustrated the sheer level of different files that could be associated with one experiment, and demonstrated a high level of organisation. Their lab book entries were very organised and formulaic with reactants, properties, quantities, densities, and safety data all marked out before going into the lab; these were the types of structured layouts that would lend themselves to using templates. They also make use of ACD labs and other software packages to visualise their data.

This was a very interesting observation as this was the first (and only) lab situation encountered where there was an actual space where computers and technology couldn't be present, although admittedly this was a small space and didn't stop the presence of computers in that room. Nonetheless it does add to the consideration of accessibility in certain areas, where scientists perceive that for multiple reasons it would be difficult to take their laptops into some lab spaces. For the crystallographers this was more based on space, whereas in these rooms this due to different types of hazardous environments for technology. These observations also emphasised, like with the crystallographers that there are so many different files and software packages that there is clearly a need for some form of unification.

4.4.4.3 Inorganic Chemists

In this research group setup, there were also multiple working environments for the students. Similarly to the molecular chemists, there was a computer room where the students had their desks, a main lab area, which was set up for conventional bench chemistry and looked similar to the molecular chemistry lab, and finally an x-ray room which was a much smaller room on a different floor with specialised equipment.

At the time of observation, there was a mix of students in the lab and at their desks. Some students were sat at their desks, Participant AA was making graphs in Excel, with data obtained from scanning materials with x-rays and retrieving the data off the computer. They were using Excel to plot the graphs and manage the data, with help from their paper lab notebook on the desk. Participant AB was also using Excel, in this instance to make complex tables of data, taking data from different data files off their computer and combining them into tables. Participant AC was using their computer to order stock but also had Word open on another screen. These activities were all related to data modelling to create material for writing up their experiments; fitting with the outcomes of the focus groups that writing reports and the associated tasks are the activities that lend themselves most to using technology. This also illustrates that the paper lab book was being used to aid in this process, suggesting both that at least some of the material from the lab book will end up digitised, but also that there could be an element of duplicated data entry, an area that concerned scientists when considering using an ELN.

Inside the main lab room, students were wandering in and out, and wearing lab coats, and performing experiments with their lab books out on the desk. This lab also had a very social atmosphere and the students clearly got on well. There were two computers in this lab that were switched off and did not look like they were in use. After enquiring about these computers, Participant AD explained that one was linked to a machine for the purposes of reading data off it, and the other was for inventory. This shows a pattern of computers existing in the lab, but only being used for specific purposes, typically to read data off a machine or in this case for organisational purposes.

When Participant AD was asked why they wouldn't want to make notes on a laptop rather than using their paper lab notebook, it lead to this conversation:

Observer: 'Why wouldn't you use a laptop as you're measuring something and then making notes in your lab notebook?'.
Participant AD: 'My laptop is the only one I have, I wouldn't want to ruin it'.
Observer: 'What if you had an excess of cheap tech and it wasn't a case of ruining one precious laptop'.
Participant AD: 'If it was cheap and durable maybe'.
Observer: 'Do you think that would be better than paper or not?'.
Participant AD: 'Paper still seems quicker but you don't know until you try'.

Participant AE extended an invitation to the x-ray room (which is shown in Figure 4.10); this room had a lot more computers in it that were actually being used. The computer Participant AE was using showed real time data on the screen and then provided the data files once the x-ray had finished. Similarly to the organic chemists, the students who use this x-ray machine often write down the values produced into their lab book, and then transfer them into excel. It was mentioned that these machines will only save the data produced as PDF's which means that they then find it difficult to get the data out of it in a readable form.

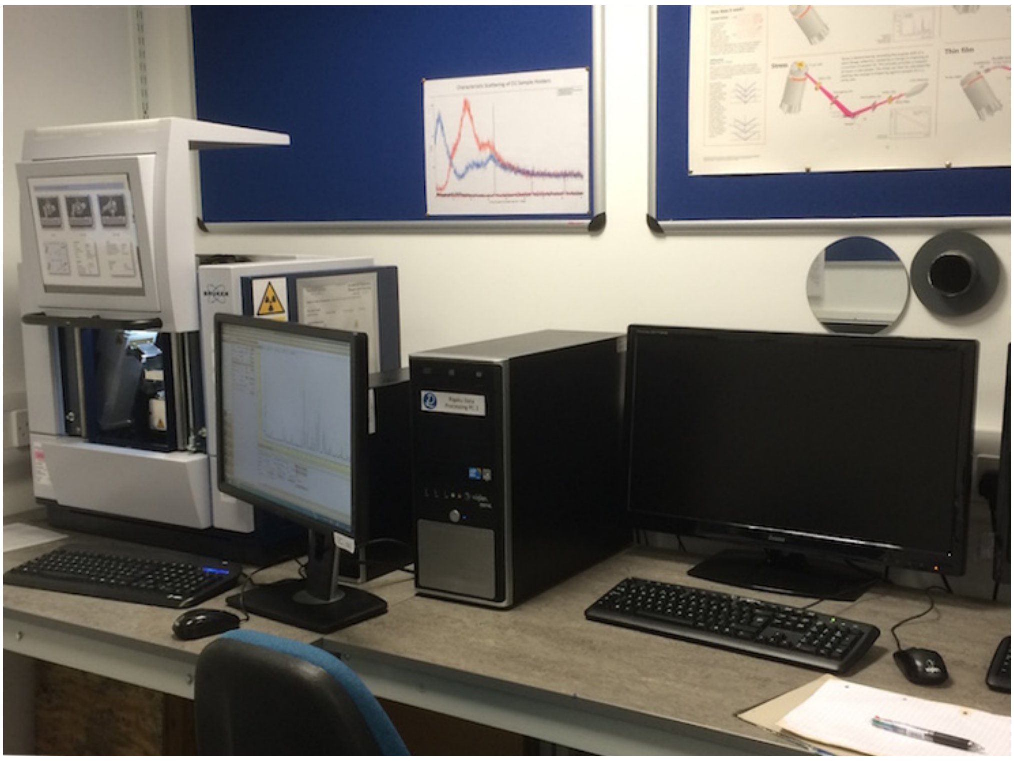


FIGURE 4.10: The XRay Room

As a stark contrast Participant AE thought that ELNs were a fantastic idea: “it puts everything in one place, there are too many different data files of different formats, for one material there could be up to 8 data files for a basic characterisation, and it gets very frustrating and it’s very hard to link together”. They then continued to explain how they write up all the important information from their lab book onto their computer. They went into more detail about their industry sponsors, how for work involved with them they are not allowed to use Skype or Dropbox as they are considered insecure, but for non-sponsored work cloud based software is used; following up with the fact that they found navigating through somebody else’s lab book very difficult and confusing. This demonstrates further that there is a need for better integration and management of the tools used by scientists.

Participant AE also mentioned that their supervisor was in favour of ELNs and had informed them about their existence. This could link with the social influence factor of UTAUT as a more important figure had expressed a positive opinion about this type of software. Furthermore, this could show that in some cases social influence could override the concerns of adding a new actor to a network, in this case the ELN.

4.4.4.4 Organic Chemists

This chemistry lab setup was similar to that of the inorganic and molecular chemists, with a computer room that housed the students’ desks, and a connected main lab room (displayed in Figure 4.11) with a conventional bench chemistry setup.

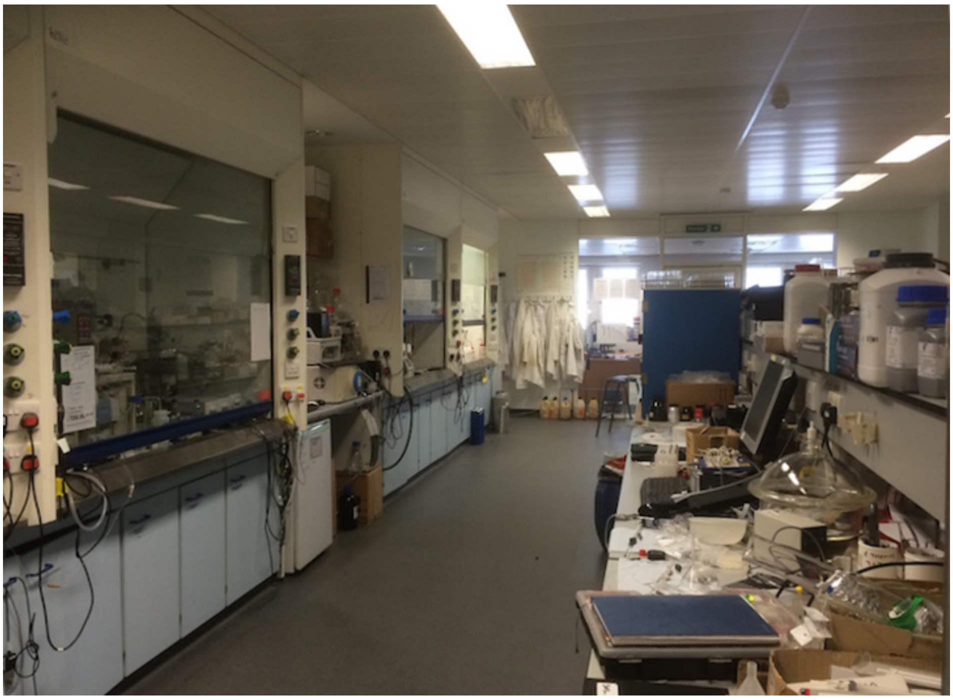


Figure 4.11: Organic Chemistry Lab

There was a friendly atmosphere in the lab, and all the students seem to get on; with some helping each other or discussing their work. Students came and went frequently, some would briefly come into the lab to change something in their experiment, or would leave the lab to make a note of something at their desk or send an email. The lab felt very paper-based, partially due to the presence of multiple paper notebooks, and also because of the paper based systems used for organisational purposes. The lab housed a paper folder that was used to store the COSHH forms, and there was a sheet of paper containing a table stuck up on the wall for people to fill in if they wanted to add more chemicals to the database; all the labels on the glass doors of the fume cupboards were written in pen that could easily be wiped off and written over. Additionally, this lab environment looked very hostile to technology, thus explaining why certain chemists had stated that they wouldn't want to take their technology into the lab. A lot of the lab coats were covered in chemical stains and all of the students were wearing gloves to touch any of the chemicals they used in their experiments, due to their hazardous nature.

This lab also contained several computers that were switched off and not currently being used during the time of observation; these computers are shown in Figure 4.12.

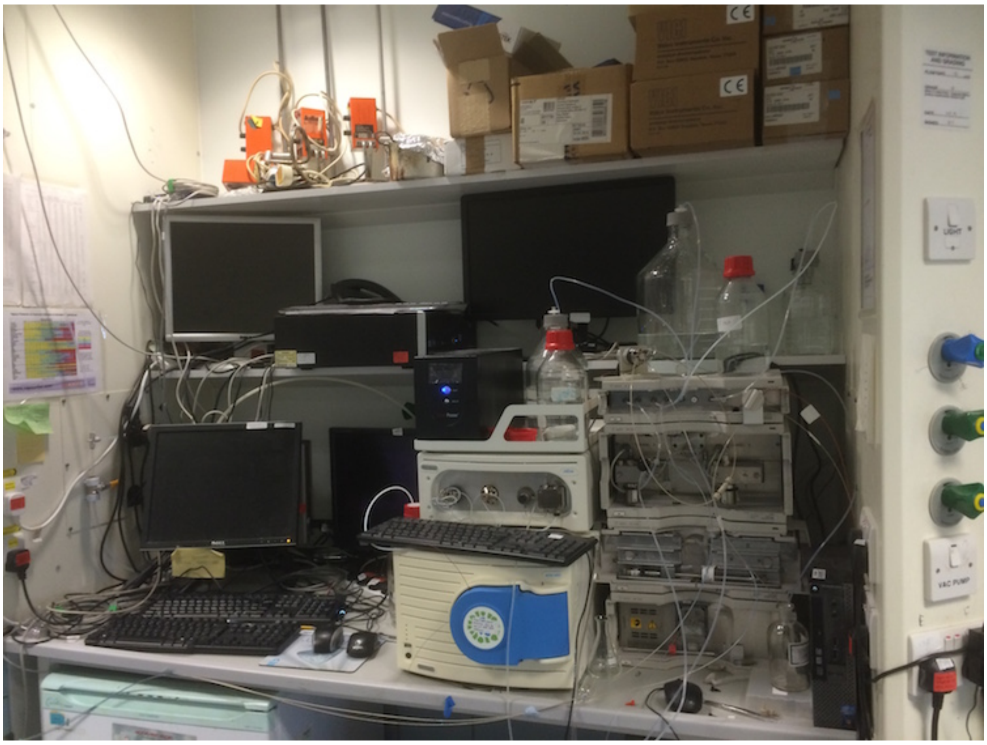


FIGURE 4.12: Computers in the Organic Chemistry Lab

Enquiries about the purposes of these computers found that one holds their chemical database, and is mostly used for searching for chemicals, although it was mentioned that it was also linked to 'Flow Chemistry' and one of the students not present in the lab at the time used that to program their reactions frequently. There was another computer where no participants in the observation (or associated students who the present participants might have spoken to) were aware of its purpose. Participant Y extended an invitation to the instrument room. The computer that the chemists used in the instrument room (a room full of computers and equipment). It was explained that this computer was used for their gas chromatography machine; it was an old Windows NT machine that had replaced a Windows XP machine that had recently stopped working. Furthermore, Participant Y explained that the students have to make notes by hand to capture the results and values generated on the machine, because that was currently the only way they could do things. This bore similarities to the crystallography observations where values had to be entered into a paper sample book and transferred between different pieces of software; illustrating that these lack of integration issues exist across the different sub disciplines.

Another student (Participant Z) had actually brought their laptop into the lab, but was merely using it for music, whilst still making notes in their lab book by hand during their experiment. This student seemed particularly resistant to the idea of ELNs and said that they much preferred to make notes by hand as and when they were doing things as part of their experiment. This could be due to the affordances of paper that still exist over technology, however it could also show that in this particular student's case, they felt that the introduction of an ELN as a new actor to their network, or indeed the loss of their paper lab notebook from it would be highly disruptive. Participant AI was less against the idea, having used one in an industrial placement, although did comment that whilst they didn't mind it, when it crashed it brought all work to a resounding halt. This also links to the facilitating conditions factor of UTAUT as if individuals do not perceive that there will be a good supporting infrastructure in place (in this case a back-up method for the ELN crashing) then they will be less likely to consider using a new piece of technology. However, their main comment about the use of ELNs was that they didn't want to change their current way of working, because *'once you've started doing something one way, you don't want to change it'*; demonstrating an unwillingness to change mid-way through a process. This suggests that if ELN adoption was to be increased then steps may need to be put in place at the beginning of a student's academic career, rather than trying to persuade them to change mid-way through.

What was very interesting was that despite not using an ELN, Participant Y showed signs of working within this environment. They took paper notes in their lab book, and then wrote them up neatly on the computer at a later date, and therefore clearly wasn't resisting an ELN because of the potential duplication of data that so many people take issue with, and indeed one could argue that in an environment where those types of notes are already performed, this would be the perfect environment to introduce an ELN for the scientists to use outside the lab.

4.5 Discussion

The results of these studies backup some of the previous findings discussed in Chapter 2, whilst also giving new insight into the current attitudes of scientists towards lab notebooks, both electronic and paper. The survey results in particular emphasise the sheer amount of different software packages and data that chemists have to deal with during their work, and make it abundantly clear that an ELN that was built specifically to cater to the needs of one type of chemist would not then work for all the other types of chemists out there, because they all work in different environments, with different software needs and all have different ways of working. It has been raised by all three user studies in different ways that there are too many different types of software and subsequent data files and that some unified systems and ways to convert these file formats could be really useful.

The results of the focus groups and participant observations suggest that the paper lab notebook is a very significant actor (as was postulated in Chapter 2), that holds a lot of attachment to the scientists in the lab. The potential removal of this actor would obviously cause huge disruption across the lab. However, despite this, these results also suggest that there is clearly a wide use of technology across note taking, just in certain areas. Popular features in scientific records as identified by Shankar (2007) such as calculations, creating graphs and figures and writing up publishable material are mostly produced using software, whereas taking every day notes, writing down ideas and noting down observations and indeed all the things that need to be done quickly with complete free editing ability is still done on paper or in a lab book.

Some participants are moving towards using software for creating their records or transitioning their notes onto a computer, but many still like the freedom paper offers; although they do seem to have imposed their own personal standardisation methods of data entry and record keeping on themselves. The participants have made it clear that they still believe that the affordances of paper outweigh the advantages of technology in certain situations; and in other situations they felt that paper was the only option. The focus group results revealed that the participants still find paper faster and easier to use when jotting down notes in the lab, or for making notes when thinking about their work or adding annotations to literature papers. Furthermore, the lab observations illustrated certain situations where some scientists couldn't just transfer pieces of information seamlessly between the different software packages that they needed; or where an old machine with limited capabilities wouldn't produce data or outputs in a format that could be used, so these outputs needed to be manually written down.

The current lab set ups that some of the chemists work with is also not conducive to using technology in the lab. There are concerns about damaging the necessary computers with chemicals, duplicate data entry and work often taking place in areas that aren't suited to a computer either because of the equipment around or space to put down laptops or even internet access. Participant S stated that 'there are no bits of tech I'd be willing to take into the lab'; which impedes the use of anything more than a paper notebook in these type of lab situations. However, that obviously doesn't extend to all chemists, and those participants have shown a willingness to use certain items of technology provided by the lab such as the cameras attached to microscopes which can then be used to get their images off, and Participant T said that they sometimes used their tablet in a separate room attached to the lab that was used for more analytical work. This could therefore be taken as less of a reluctance to use technology, but a reluctance to use technology in a situation where it could get damaged, or to fully give up their paper lab notebooks. There is also obviously a need for certain updated software solutions, when the archaic procedures of copying down information from a computer by hand or manually comparing spreadsheets are still used.

The current hostile lab environments, combined with both the affordances of paper that still outweigh technology, and some strong reactions at the notion of either adding an ELN or removing the paper lab notebook, suggest that an incremental solution may work better than trying to directly replace paper. It would be more sensible to produce an environment that will work with a system that still uses paper. The results of Shankar (2004)'s studies concluded that the scientists still worked with a large amount of paper and that doesn't seem to have changed. It logically follows that any viable ELN environment to work in these conditions could not hope to entirely replace paper as there will be some people or some circumstances where they will not want to have a laptop or other appropriate piece of technology with them at a crucial moment; alongside the fact that investigations of fully electronic systems such as (Hughes et al., 2004a) have proven to be overly time consuming, which puts users off. Therefore, any such environment will not only need to be created to work in harmony with the paper lab book, but also to link up with the other technological devices that might be used such as smartphones, and be available cross platform (Tabard et al., 2008). Perhaps an environment that was also available on a smartphone or facilitated easy linking between photos on a smartphone and the platform would provide some of the necessary bridging between use of tech in the lab, paper in the lab, and a notebook environment. Additionally, all three studies have elicited results that highlight issues with the current software offerings available for scientists, software packages do not integrate well, and there are arguably too many different tools for a scientist to choose between. An ELN should look to integrate existing software both from a domain and note booking perspective that scientists already use to provide a better integrated system that can be used for more than one activity.

Furthermore, as discussed in Section 3.1.1 the ELN also needs to be looked at as a separate actor and it is treated as such throughout these studies. Whilst other software packages are used and discussed by chemists without too much of a second thought, and just seem part of everyday lab life, the concept of an ELN seems to be much more serious and closer to the hearts of the scientists. The works of (Shankar, 2004) denoted that the creation of these lab records are a very personal endeavour, and the results of these studies lend credence to that theory, given the contrasting ways in which the different scientists organise, create and index their notes, in addition to the concerns voiced that an ELN would not be as customisable as a paper lab notebook.

These scientists are reacting to the potential of a new actor being introduced to their lab network, and whether they are supportive of the idea or not, it is clear that none of them have a neutral reaction to it. Therefore, when considering creating and introducing this new ELN platform, it must be treated as a new actor, and it's effect on the entire network must be considered. A locked down ELN environment where there wasn't much freedom for how to organise the notes wouldn't be attractive to most scientists. It is also influenced not only by the different subjects they are studying, but also by how their peers / superiors / group members organise their work, so any ELN environment would require a high level of flexibility with regards to how it can be organised. Additionally, there are hints that even by those who wouldn't necessarily classify themselves as using software or platforms to take their notes, there is an underlying element of technology usage.

With respect to UTAUT, the study participants have made it clear that ELNs are not perceived as either easy to use or useful by some scientists. A lot of the descriptions given of what participants would like to see from such platforms, or even what they desired from software packages in general was programs that made their life easier. The scenarios presented in the focus groups suggested that the ease of use of paper outweighed the potential loss of lab work; illustrating that the Effort Expectancy factor is a highly valued one. It seemed that the participants place a higher value on their electronic data rather than their paper work. Of equal importance is the Performance Expectancy, as various participants who are against the idea of ELNs perceive them as not being useful enough to be worth the effort, and those who want to use them believe them to be useful tools. Furthermore, some social influence elements have become clear throughout this study. It became obvious which participants had supervisors who believed that ELNs were a good idea, as they were much more in favour the idea of one even if they didn't currently use one. Additionally, facilitating conditions also seemed to be an influencing factor to some participants in these studies, as there were some participants who were using software and systems that didn't work very well, including a participant who noted that the ELN they had to use in industry wasn't too bad apart from when it crashed, which meant that all work had to stop until it restarted again. Scientists may not be willing to use an ELN or indeed any software instead of their paper lab notebook if they are concerned that it won't work in the way they want it, or if they can't use it for a period of time if a crash occurs.

Consequently, perhaps any ELN environment needs to be geared towards the storage of the data they already record electronically, whilst also facilitating the sort of notes that will be made in paper. In addition, any ELN created must both be easy to use, and provide useful services to its users, and in this case, all the different types of chemists. It is vitally important that the ELN environment is presented as a platform that can make the scientists life easier and work with their paper lab books to whatever extent they desire, as opposed to insisting that it is a full replacement for their paper notebooks. (Shankar, 2007) also suggests that these lab notes are only viewed as useful in the short term, with (Latour and Woolgar, 2013) finding that scientists often only write up their work outside the lab book for publication. Therefore, any ELN environment would not only need to fit within that way of working, but prove useful enough and simple enough to use that it would be worth using at some point in the experiment life cycle, either as a tool for creating the final published document, or an area to store or create notes if the user desired.

Interestingly, despite the concern that using ELNs would lead to duplication of work, several of the participants of these groups already essentially go through this duplication process as they have to write up and or reformat their data to share it with people outside the university, or in some cases their supervisor. They also all create presentations, reports and publications out of their written notes and electronic notes. Thus, it would be important to make an ELN environment that fitted in with this culture, and that facilitated these procedures. It would also be prudent to point out that there is clearly an element of duplication of data entry, so if students are writing notes by hand and then writing them on a computer, why not use an ELN environment.

Overall it seems like there is a place for an ELN environment that serves as an interface between the paper lab notebook and the electronically created notes. Following the principles of ANT it is clear that attempting to remove the paper lab notebook would cause immense disruption, and there are varying levels of opposition to this idea. It may help to reduce the levels of equal opposition to the introduction of the new ELN actor if it is brought in alongside the paper notebook. Additionally, following the ideals of UTAUT any ELN environment must be easy to use and useful, as paper still holds a high place in the lab system due to its ease of use and flexibility, and that in order to be persuaded to change their current methods any ELN environment would have to not only be perceived as useful, but prove to be more useful than their current systems or there would be no incentive to change.

4.6 Key Findings

The focus groups, in particular with support from both the survey and the lab observations, provided valuable insight into the current ways in which the scientists work in their labs and what types of functionality they were looking for from an ELN; the key findings will be summarised in this section.

4.6.1 KF1: How scientists take notes and organise their work is a highly personal endeavour

The focus groups in particular have made it clear that scientists take notes and organise their work in very different ways. Whilst there were some common pieces of information recorded such as dates, and links to what experiment or project some work is part of, overall the scientists recorded different pieces of information such as sample values, weights, measurements, and temperatures. Some scientists had very structured organisational processes, with the chemists even using custom made codes to provide links between paper and electronic data so that a pathway can be followed if they are trying to write up some work. Furthermore, even the ways the scientists organised and created their electronic notes also differed vastly. Some participants had highly organised complex file hierarchies, whereas others were more sporadic with how files were labelled and where they were stored. Shankar's 2007 study concluded that note taking was a very personal endeavour, and ten years later this doesn't appear to have changed. Oleksik et al.'s 2014 study concluded that even using the same piece of software, people created notes very differently. This diversity illustrates that whilst functionality such as standardised templates could be useful for speeding up data entry, it wouldn't be attractive to the users to restrict how they could create and organise their notes, as they would lose their personal way of managing their work. Some participants in the biology focus group noted that they like being able to customise their notes with colours and highlighting, and noted that writing up their work was less personal and felt more formal. In order to entice scientists to digitise their notes as well as formal written pieces of work, such customisation and flexibility would need to be offered.

4.6.2 KF2: Scientists do use technology to digitise portions of their work despite their continued use of paper

All of the participants from the focus groups agreed that they used electronic devices to write up their reports, posters, publications, and indeed any formal documents that were required of them. The lab observations reinforced this, as the participants who were in the lab were conducting experiments and using their paper lab notebooks for notes, or using the computers in their respective labs for the purpose of reading data from a machine; however almost all of the students sat at their desks were writing reports or creating figures from their data to go in their reports. Therefore it is important to note that just because there is a resistance to the idea of using ELNs, this doesn't mean that scientists do not use technology to write up portions of their work. Additionally the software survey illustrated that chemists do use many different software packages to create structures and diagrams and to aid in the work they do. However, as illustrated by all three studies there are flaws in some of the software packages and they do not always work together well. Therefore whilst it is important to try and improve the digitisation of the scientific record, there is also work to be done to improve the current digitisation and management processes, as perhaps if the current software tasks weren't so onerous then more scientists would be willing to digitise their records further. This links strongly to the Effort Expectancy factor of UTAUT, as the scientists would need to feel that software is easy to use before being willing to adopt it.

4.6.3 KF3: Some of the elements of the scientific record that do not get digitised are the things that do not work

When asked in the focus groups about what the scientists would lose if their labs were set on fire, there was a conflict between some scientists who thought that they would lose a lot of work and that it would be very damaging, to others who didn't think that the material in their paper lab notebooks that didn't exist on a computer would be much of a loss. Some of the participants stated that the elements in their lab book that didn't make it onto a computer were the parts of experiments that didn't work, or alterations to methods and protocols that didn't have the desired results; suggesting that they wrote up most of their paper notes for working experiments, and that the failed methods were left in the paper lab book. However, it's worth noting that knowing which approaches do not work can be vitally important, both for the person who initially tried them in case they later forget and attempt them again, and particularly for students continuing or referencing their work at a later date. The comments from the focus group participants about how much their supervisors or colleagues would struggle to access their work if they were indisposed, combined with the general lack of paper lab notebook backup, suggests that not all of the participants have been considering the longevity of their work, and are currently only digitising what is useful to them with respect to formalised write-ups. Additionally, the participants who have made backups of these paper notes to ensure that later students could potentially view them did so at the request of their supervisor. There may need to be some changes in behaviour and organisation of students, potentially using social influence to establish sensible practices at the beginning of their academic career to improve the digitisation and backup of all their work, not just the successful pieces of it that end up in published papers and reports.

The participants made it clear that the software packages or technologies they used that didn't work well weren't used out of choice, an old machine was used because it was the only computer connected to a machine that scientists needed to obtain data from, and the newer computer that had been there was broken. Badly integrated software packages were used because there currently wasn't an alternative to using multiple pieces of software to compute different elements of crystal data, and the crystallographers noted that they were hoping for a better functioning and more integrated piece of software soon. Whereas the software the scientists used out of choice were the programs that they felt actually added to their work, ChemDraw was mentioned as a useful piece of domain based software, and for more generic software Word and Excel were frequently mentioned, as these are popular, easy to use, pieces of software. This links to the Performance Expectancy element of UTAUT as the software the scientists used by choice were packages they believed would improve their performance, such as writing up a neat report in word as opposed to trying to hand write it. However, in instances where they thought paper was quicker and easier and would provide a better performance than using an electronic device, such as making quick notes in the lab, they used their paper lab notebooks instead. An ELN therefore should also focus on trying to improve their current software offerings rather than create another new piece of unconnected software, especially one that aims to replace an actor as significant as the paper lab notebook.

4.6.4 KF4: Scientists are much more concerned with backing up electronic work than paper based work

When discussing backups with the focus group participants, there was a high disparity between the efforts that were made to backup electronic work versus their paper-based work. Every single participant backed up their electronic work regularly, some in multiple locations. As a stark contrast, most of the participants didn't backup their paper lab notebook material at all. Some participants used special lab books with carbon pages that would automatically create copies of their paper notes, but in many cases they didn't even rip out the carbon pages, or store them somewhere different, meaning that in the event of a fire they would be destroyed alongside the original paper lab books. This could be partly due to a learned behaviour that students are impressed upon to backup their electronic work, or that they valued their electronic work much more highly than their paper based work. In some instances this is unsurprising, as part of their electronic work are important thesis documents and papers that they wish to submit for publication. However, this also poses a problem as it shows that not only do some parts of the scientific record fail to get digitised, but that they are also not well protected against loss.

4.6.5 KF5: Many lab setups aren't conducive to using electronic devices

The lab observations showed that there were several situations that were not conducive to using an ELN. There were some lab environments that were very hostile to technology, either because hazardous chemicals might get spilt on them, or magnets might wipe their hard drives, or even in some instances there wasn't space to put a laptop down somewhere safe. This shows that there is another significant factor towards wanting to still use paper in the lab in these circumstances, in addition to the general affordances of paper for quick and easy note taking. Further to this, if there are situations where scientists would be unwilling or unable to take their electronic devices in the lab irrespective of how good the software was, then the lack of ELN adoption cannot be purely considered as a software problem. This also raises other issues such as duplicated data entry which was detailed in the concerns regarding ELNs in Chapter 3, as if the scientists cannot physically enter their work into an ELN during their experiments, then all these notes would need to be transferred to an electronic device at a later data, increasing the workload. Furthermore, it seems unlikely in these scenarios currently that scientists would be willing to even trial an ELN based on the hardware barriers. If an ELN was to replace paper, then this issue would need to be addressed, potentially in some cases with cheaper more durable hardware, or better practices where the scientists could still use the devices without being as concerned about them. Currently, without these procedures in place, there is a limitation to how much an ELN could be fully used throughout a lab, therefore it might be worth first addressing the other software needs first and improving the current digitisation processes before trying to phase paper out.

4.6.6 KF6: Scientists seemed more in favour of ELNs if they had supervisors who thought they were a good idea or encouraged their usage

The Social Influence factor of UTAUT can be seen from talking to some of the participants. Some participants who had supervisors who were in favour of ELNs also thought that they were a good idea, and particularly an observed postgraduate student who was close to finishing their project expressed desire to have used one from the beginning. Other participants who had heard of ELNs from either their supervisor or another lecturer in the department had trialled LabTrove, an ELN that had been created by the University of Southampton; suggesting that they had been more inclined to try a tool that had been specifically recommended to them by someone more important. Additionally, the other instances where scientists had used ELNs had been in industry, where they had not been given the choice, rather they were told by their boss or manager to use one (another more important figure). Similarly, some of the students who had better backup procedures in place for their paper lab notebooks were those with supervisors who had insisted on more stringent measures than the students who were left to their own devices. This suggests that social influence should also be considered with respect to encouraging students to use ELNs.

4.6.7 KF7: Scientists still seem very attached to their paper notebooks, and do not believe that an ELN would fit into their current workflow in a way that would make things easier

Despite the advantages of storing notes electronically such as easier searching and improved backup capabilities, the scientists remain highly attached to their paper lab notebook for one simple reason, they find various note taking activities much quicker and easier to conduct on paper. This illustrates that the paper lab notebook is still an important actor that the scientists value for its advantages. Many participants stated that when performing their experiments, they found it easier to quickly jot down values and observations in their paper lab notebook, which they were less concerned about moving around or putting down near hazardous chemicals. Some of the participants in the lab observations stated that they already had a method that worked and they would be unwilling to change it, or that they felt it would take more time to make their notes on a computer rather than quickly using paper. This strongly links to the Performance Expectancy element of UTAUT as the participants didn't expect that using an ELN would improve their performance. One participant stated that they didn't think an ELN would fit in to their workflow in a way that made their life easier, also linking to the Effort Expectancy element of UTAUT.

In addition to inside the lab, other tasks also seemed to lend themselves to paper, for example, many participants said that they liked to print out their literature papers and annotate them by hand, as they found this easier than performing these actions on a computer, and that they found it easier to read a physical copy of the paper. The physicists in particular said that they used paper to write out their equations as this was much quicker and easier to do than trying to input all of the symbols into an electronic program. There are clearly areas that paper still outweighs electronic devices for speed and ease of use, and the overall feelings towards how well electronic devices can perform certain tasks compared to paper would need to be improved and addressed before scientists would be willing to change their usage of paper in certain situations.

4.6.8 KF8: The software needs identified by scientists weren't necessarily what they would attribute to an ELN, rather than an improvement to their current software offerings

When the survey participants were asked what software they would like to see created, they didn't say Electronic Lab Notebooks. In some instances, they didn't even suggest a specific type of software, they said that they would like the current software offerings to integrate better with other packages they need to be used in conjunction with, or that would like data formats to either be easily converted between or normalised across different packages. The lab observations elicited many areas where there were poor software offerings, with packages that didn't integrate well together, and other programs linked to machines that only produced data in unhelpful formats, or even only on the screen such that it had to be manually written down into a paper lab notebook. When discussing software needs rather than an ELN, scientists seemed a lot less opposed to the idea of general software improvement and mentioned areas where this could be made better. This strengthens the argument that the paper lab notebook is a significant actor that would cause disruption if it was removed from the network, and similarly that the electronic lab notebook would cause disruption if it was added to displace the paper lab notebook. Furthermore, this shows that potentially be marketing the idea of an ELN differently, by proposing it as a further aid that enabled write-ups and domain based functionality, and could also be used as a note taking tool, rather than suggesting it should be used as a paper replacement could prove more popular.

4.7 Initial Conclusions

The user studies have made one thing very clear, that we cannot hope to fully replace the paper lab notebook anytime soon. Until we have the technology where a screen can be written on as accurately and easily as paper, and unless all labs are fully stocked with cheap, durable and easily replaceable tech to use instead of paper there will always be tasks that require the traditional lab book. However, despite the preference some scientists show for their paper lab notebooks, the participants of the focus groups and lab observations all have varying degrees of their work stored electronically. Therefore this project looks to create an ELN environment that can serve as an interface between the paper lab notebooks and the electronic documents that the scientists create, and that utilises modern web technologies such as the semantic web and the cloud.

Plenty of the scientists use generic note booking software such as Word and Evernote to write up their notes, Excel to handle their figures and create graphs, and their data tends to be stored electronically in different places, with different pieces of speciality software used to aid in their work. Additionally, some students have progressed to using their computers and these pieces of software to take some of their everyday lab notes as well, or at least neatly write them up into there. There is a wide usage of cloud based software and methods to make their work not only secure and backed up, but available in multiple different places on different machines. Therefore it seems sensible to build on this with the required domain knowledge rather that create a new system from scratch. Additionally scientists have identified being able to search through their data better as a high priority, and the desired features that have been elicited from the surveys include areas like storing metadata and tagging/classifying notes which can be achieved using semantic web technologies.

There are clearly areas which scientists feel could be improved with the current software offerings, even if they do not necessarily want to use an ELN. This project is aiming to improve the digitisation and management of scientific records, and so is looking to conceptualise a piece of software that reflects the key findings of these studies and fulfils the user required features detailed in Figure 4.3 to see if that would have a positive impact towards this goal. This project looks to create a piece of software that provides a centralised location for the scientists to manage all heir notes and data in the cloud that also has semantic capabilities. This software will also serve as a traditional ELN in that it could be able to be used in the lab if the scientists so require, but equally is not looking to necessarily replace the paper lab notebook in instances where this would not be practical, rather to offer better services to manage and digitise scientific records, in the hope that this would encourage further digitisation and better management.

Chapter 5

A Semantic Cloud Based ELN

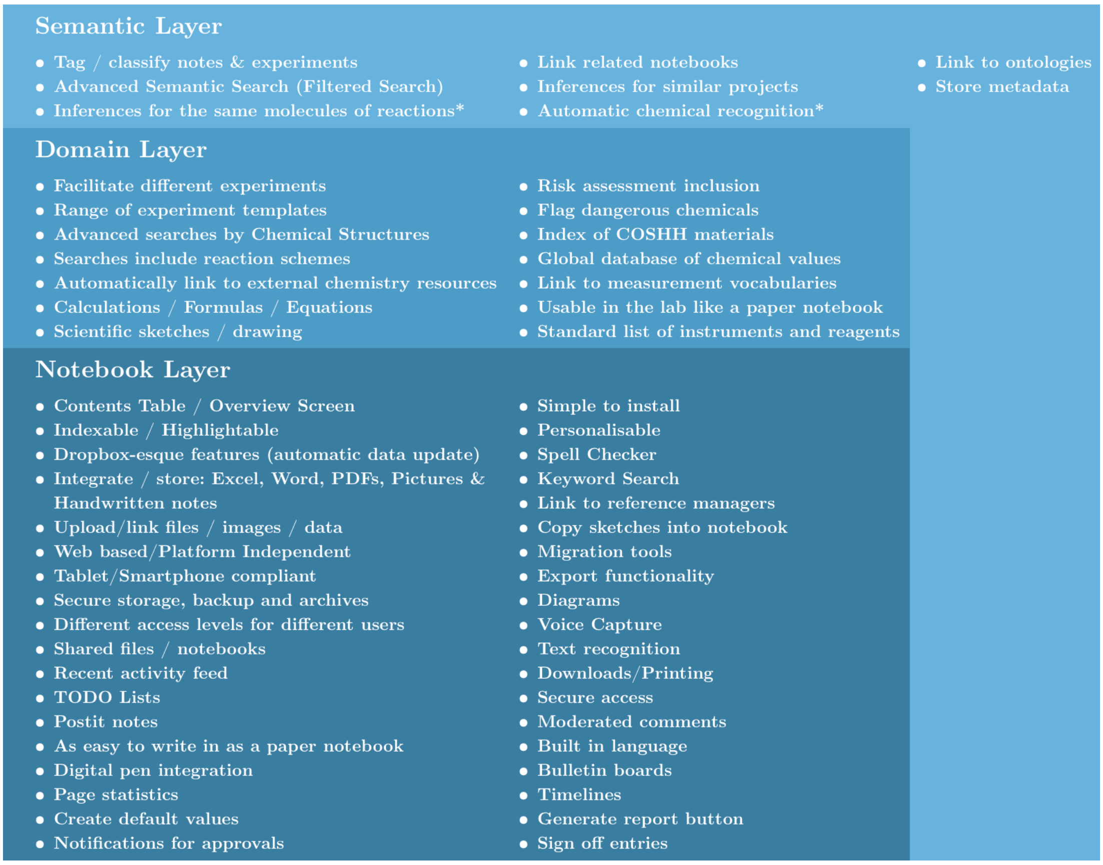
This chapter will cover the main technical aspects of this project. It will detail the design and conceptualisation of the entire proof of concept ELN and the development of Semanti-Cat, the semantic layer prototype that was created as part of this project. The proposed requirements of the overall system will be laid out, followed by a review of the different cloud notebook platforms that this system could be built on top of, explaining which platform has been chosen for this project and why. The next section will detail some early experimental coding with the chosen cloud platform to understand how some of the requirements of the domain layer could be implemented, and to see how 3rd party tools can be integrated with it for both the domain and semantic layer. Following this, the development of Semanti-Cat will be explained (as given the time and scope of this project only this layer was created), detailing which features have been implemented in this first iteration and why. A full breakdown of each part of the system will be given, followed by an explanation of the tools and technologies used in this prototype. This chapter will finish with a walkthrough of Semanti-Cat, with screenshots and descriptions of the functionality that each section provides.

5.1 System Design and Requirements

To formulate the design and requirements for this system, the user desired features detailed in Table 4.2 have been broken down into three layers:

- Notebook Layer - contains all the features that are notebook specific and could potentially be found in a generic electronic notebook.
- Domain Layer - contains domain specific features (in this instance with a specific focus towards chemists but could be replaced with any other domain in this design).
- Semantic Layer - contains all the metadata / additional semantically added information).

This three layered design was conceptualised in the original model (Figure 2.2) that illustrated how the features of a semantic notebook incorporated both an electronic lab notebook and an electronic notebook, and similarly how an electronic lab notebook incorporated the features of an electronic notebook. The features detailed in Figure 2.2 were generic features taken from literature. These have been replaced with the user desired features from the initial user studies and the Dial-a-Molecule studies that were analysed as part of this thesis (from Table 4.2), which have been categorised into the layers of the original model.



Semantic Layer

- Tag / classify notes & experiments
- Advanced Semantic Search (Filtered Search)
- Inferences for the same molecules of reactions*
- Link related notebooks
- Inferences for similar projects
- Automatic chemical recognition*
- Link to ontologies
- Store metadata

Domain Layer

- Facilitate different experiments
- Range of experiment templates
- Advanced searches by Chemical Structures
- Searches include reaction schemes
- Automatically link to external chemistry resources
- Calculations / Formulas / Equations
- Scientific sketches / drawing
- Risk assessment inclusion
- Flag dangerous chemicals
- Index of COSHH materials
- Global database of chemical values
- Link to measurement vocabularies
- Usable in the lab like a paper notebook
- Standard list of instruments and reagents

Notebook Layer

- Contents Table / Overview Screen
- Indexable / Highlightable
- Dropbox-esque features (automatic data update)
- Integrate / store: Excel, Word, PDFs, Pictures & Handwritten notes
- Upload/link files / images / data
- Web based/Platform Independent
- Tablet/Smartphone compliant
- Secure storage, backup and archives
- Different access levels for different users
- Shared files / notebooks
- Recent activity feed
- TODO Lists
- Postit notes
- As easy to write in as a paper notebook
- Digital pen integration
- Page statistics
- Create default values
- Notifications for approvals
- Simple to install
- Personalisable
- Spell Checker
- Keyword Search
- Link to reference managers
- Copy sketches into notebook
- Migration tools
- Export functionality
- Diagrams
- Voice Capture
- Text recognition
- Downloads/Printing
- Secure access
- Moderated comments
- Built in language
- Bulletin boards
- Timelines
- Generate report button
- Sign off entries

FIGURE 5.1: Revised Feature Model, adapted from (Kanza et al., 2017)

Figure 5.1 illustrates both the three layered design of the system, and also the required features that exist in each. Splitting up the layers means that they could be worked on separately and packaged on a standalone basis to provide a set of chemistry tools that could be integrated with a notebook, or some semantic software that will automatically tag scientific documents. Splitting out the domain layer means that this design could therefore work with any different domain, the chemistry specific elements could just as easily be replaced with biology-based features for example. The semantic layer could be customised based on the domain layer, with a base set of features to tag and classify the notes that use different ontologies and knowledgebases depending on the chosen domain. Additionally, separating out the notebook layer means that there the possibility to find a cloud notebook that possesses most of these features already such that the other layers can be built on top of it. The two starred entries in the semantic layer represent the two features that are heavily domain specific but that still belong in the semantic layer.

The functional and non-functional requirements that have been identified from Figure 5.1. The non-functional requirements are those that specify how the system should work as opposed to what specific functionality it should contain, whereas the functional requirements denote what the system should do. These requirements are listed as follows:

5.1.1 Non-Functional Requirements

NFR1: The system must be secure and facilitate different access levels

NFR2: The system must be simple to install

NFR3: The system must be as easy to write in as a paper notebook

NFR4: The system must be usable in the lab like a paper notebook

NFR5: The system must be customisable

NFR6: The system must be usable across all devices including computers with any operating system, and smartphones / tablets

NFR7: The system must be cloud based

NFR8: The system must be collaborative

5.1.2 Functional Requirements Notebook Layer

FRN1: The system must facilitate file uploads and downloads

FRN2: The system must facilitate automatic syncing and updating between devices

FRN3: The system must integrate with common data formats (be able to handle excel/word/pdfs/pictures) with respect to file storage, file conversion and file exports

FRN4: The system must provide table of contents, indexing and highlighting capabilities

FRN5: The system must include a spell checker

FRN6: The system must provide a keyword search

FRN7: The system must provide tools to create diagrams electronically

FRN8: The system must have text recognition and drawing capabilities

FRN9: The system must be able to link to a reference manager

FRN10: The system must be able to display a user's recent activity

FRN11: The system must be able to create TODO lists

FRN12: The system must be able to create post-it notes

FRN13: The system must be able to print the users documents

FRN14: The system must be able to handle voice capture

FRN15: The system must be able to facilitate moderation of users comments

FRN16: The system must contain different built in languages

FRN17: The system must be able to be input to by a digital pen

FRN18: The system must be able to create timelines

FRN19: The system must be able to facilitate bulletin/message boards for users

FRN20: The system must be able to provide page statistics

FRN21: The system must provide a 'Generate Report' button to generate a publication ready report

FRN22: The system must facilitate notifications about entries requiring approval, and enable them to be signed off / made non editable

5.1.3 Functional Requirements Domain Layer

FRD1: The system must provide a range of templates to facilitate different experiments

FRD2: The system must provide advanced searching to search by chemical structure

FRD3: The system must include reaction schemes in the search results

FRD4: The system must link to external chemical resources including lists of instruments and reagents, measurement vocabularies, databases of chemical values

FRD5: The system must be able to facilitate scientific drawings

FRD6: The system must be able to perform calculations and use formulas

FRD7: The system must be able to handle equations

FRD8: The system must include risk assessment features, with an index of COSHH materials and the ability to flag dangerous chemicals

5.1.4 Functional Requirements Semantic Layer

FRS1: The system must tag/classify documents

FRS2: The system must store metadata about the documents

FRS3: The system must link to ontologies

FRS4: The system must provide an advanced semantic search

FRS5: The system must provide linking of related notebooks

FRS6: The system must provide inferences about similar projects

FRS7: The system must be able to automatically identify chemicals

FRS8: The system must provide inferences about projects that are working on similar molecules (which requires an inbuilt understanding of molecules).

5.2 Choosing a Cloud Based Service

There are a number of cloud based services that could be used as a platform for this project. In order to select the best option to represent the notebook layer in the proof of concept system, the following criteria were used. Firstly, a number of cloud notebook platforms were identified, and they were evaluated against the objective non-functional requirements (for example, the system being cloud based and collaborative is something that can be objectively measured, whereas the system must be usable in the lab like a paper notebook is subjective to an extent to the user). The additional criteria it was evaluated against was whether it was possible to externally develop on top of these systems, and whether it was free to use to mitigate the most significant cost barrier identified in Chapter 2.

For those notebooks that met all six categories, they were compared to see how many of the functional notebook requirements they fulfilled. The selected notebook platform was the one that met these requirements the best. The initial set of cloud notebooks to be considered included: Evernote[1], OneNote[2] and Google Drive[3]. These were considered because all three had been trialled as an ELN in previous research (Walsh and Cho, 2012; Guerrero et al., 2016; Weibel, 2016). Also included were Microsoft's Office 365[4], Apple's iCloud[5] as these were similar well-known note booking software products. Table 5.1 details a comparison of these notebooks based on the initial criteria detailed above.

Cloud Notebook Platform / Criteria	Free to use	Facilitates Development	Secure with different access levels (NFR1)	Accessible on all platforms (NFR6)	Cloud Based (NFR7)	Collaborative (NFR8)
Apple's iCloud	✗	✓	✗	✗	✓	✗
Evernote	✓	✓	✓	✓	✓	✓
Google Drive	✓	✓	✓	✓	✓	✓
Microsoft Office 365	✗	✓	✓	✓	✓	✓
OneNote	✓	✓	✓	✓	✓	✓

Table 5.1: Initial Comparison of Cloud Notebook Platforms for suitability for the prototype

[1] https://evernote.com/

[2] https://www.onenote.com/

[3] https://www.google.co.uk/drive/

[4] https://products.office.com/en-gb/compare-microsoft-office-products

[5] https://www.icloud.com/

As demonstrated by Table 5.1 Evernote, Google Drive and OneNote all meet the first set of criteria. These three platforms were then evaluated against the non-functional and functional notebook requirements to see which one met the most requirements. Google Drive met 18/23 of these requirements, with OneNote meeting 14/23 and Evernote meeting 10/23. Therefore, Google Drive was chosen as the cloud platform for the proof of concept project. The full table detailing this evaluation can be found in Appendix E.

5.3 Development Environment

Before creating the proof of concept prototype, some experimental coding was done using Google Drive's development environment. Google Drive offers two different ways of developing on top of their platform: Google App Scripts[6] and Google Drive API[7].

Google App Scripts is a JavaScript cloud scripting language that works with the Google Object Library to enable direct interaction with Google Docs and Sheets. The most relevant features for this project are: developers can create custom menu bars, custom sidebars and custom dialog boxes for Google Docs and Sheets. Add-ons can also be created to extend the functionality of Google Docs and Sheets, that can be published to the add-on store. However, these add-ons are still limited in what they can change in the document itself. For example, even when creating an add-on, a developer cannot access the right click menu or enable hovers in the document. Additionally, Google's security policy can often cause cross-site scripting errors. For example, when trialling the integration of a ChemSpider plugin widget within a Google Document (to trial the integration of third party applications for the domain based layer as part of the experimental coding), cross-site scripting errors occurred typing to link to external JavaScript files that were hosted elsewhere. All scripts needed to be moved to a folder in Google Drive and published to the web, and then hot linked to within the Google App Script project files. Additionally, Google App Scripts facilitates creating HTML files, but separate JavaScript and CSS files cannot be made inside the scripting projects, they can only be embedded in the HTML files which can lead to the long messy unreadable files. The development capabilities described here were tested out with some experimental coding, the full details of which can be found in Section F.1 in Appendix F.

The initial experimental coding suggested that Google App Scripts would work well for visualising the front-end of the semantic layer and illustrating how an ELN could be built on top of Google Docs. However, a project decision was made to create the backend of the semantic layer as a standalone web application that obtained the files to tag using the Google Drive API. This was decided for three reasons. Firstly, creating this part separately would mean that it could be packaged as a separate application/service that

[6] `https://www.google.com/script/start/`

[7] `https://developers.google.com/drive/`

could be used by or applied as an add-on to any other electronic note booking software, rather than being completely built into one application. Secondly, despite Google Drive fulfilling more of the notebook based functional requirements than any of the other systems, some participants raised concerns about the security of using Google Drive, thus it seemed more practical to make clear to participants in the software evaluation focus groups that this concept of ELN would not necessarily need to use Google Drive. Lastly, this is a very initial prototype so creating it as a standalone web application would provide more freedom from a development perspective, some limitations of Google App Scripts had already been identified from the experimental coding and given that creating the semantic layer would require the use of external semantic services, it was more practical to use an environment that didn't pose restrictions on the use of external services. The next section will detail how this semantic layer was created.

5.4 Developing the Semantic Layer

This section will detail the development of the semantic layer, including what features have been added, what tools and technologies were used and finish with some screen shots of the created application.

5.4.1 Implemented Requirements

Firstly, the functional requirements from the semantic layer were looked at to see which were the most feasible to add into this prototype given the time and scope of the project, and to see if there were any dependencies between requirements that would necessitate certain functionality being implemented first.

FRS1, FRS2, FRS3 and FRS7 were identified as requirements that didn't have any dependencies and could be implemented (tagging documents / storing metadata / linking to ontologies and automatically recognising chemicals). FRS4 is a very important requirement as improved search was listed as one of the top priorities by users for an ELN, however this relies on making sure that the documents are tagged well and linked to the correct ontologies, so has a dependency on requirements FRS1, FRS2 and FSR3 and is therefore considered out of scope for this prototype. FSR5 and FRS6 are similarly considered out of scope because in order to link related notebooks and infer about related projects they also depend on FRS1 and FSR3 to have suitably tagged and marked up the documents in order to ascertain if they are linked or not, and to understand how a user would perceive a notebook or project to be perceived to be related. Finally, FRS8 is also considered out of scope because this is very specific to one type of chemistry and this proof of concept project will be trialled with scientists from all three major disciplines. Based on these dependencies the requirements were prioritised using MOSCOW

(Van Vliet et al., 1993). The four requirements that didn't have any dependencies listed as the must have requirements, and FRS4 was listed as a should have requirement, as that could still be coded and tested on smaller scale systems, whereas the requirements that need multiple notebooks and projects to link together and make inferences about require a larger scale set of data to test on, therefore these have been listed as could have requirements.

5.4.2 Must Have Requirements

M1: The system must tag/classify documents (FRS1)
M2: The system must store metadata about the documents (FRS2)
M3: The system must link to ontologies (FRS3)
M4: The system must be able to automatically identify chemicals (FRS7)

5.4.3 Should Have Requirements

S1: The system must provide an advanced semantic search (FRS4)

5.4.4 Could Have Requirements

C1: The system must provide linking of related notebooks
C2: The system must provide inferences about similar projects
C3: The system must provide inferences about projects that are working on similar molecules (which requires an inbuilt understanding of molecules).

How these must have requirements will be implemented will be discussed in the software evaluation focus groups, and further questions will be asked to understand what users want from a search, and how they want their projects to link together.

5.4.5 Tools & Technologies

The main tools and technologies used in this system were tagging services to tag and classify the documents and thus store metadata about them, ontologies to also help with tagging the documents (and lead up to helping create the search in future developments), and services to automatically recognise chemicals. The initial coding investigations for this prototype involved using a system called LOOMP (Luczak-Rösch and Heese, 2009) which enabled users to mark-up their own documents by looking for terms and phrases to see if they appeared in one of the ontologies loaded into the system, and then linking the ontology descriptions to those terms. However, this required a high level of user input, and given that one of the adoption barriers to ELNs was the potential of additional/duplicated work, a search was conducted to find more automated tagging systems that would reduce the required level of user input to facilitate these tags. Screenshots and explanations of the prototyping work conducted using LOOMP can be found in Section F.2 in Appendix F.

5.4.5.1 Tagging Services

When searching for semantic tagging services there weren't a great deal of available tools out there specifically for tagging even if there were a decent amount of research papers about the topic. The two identified services in this search were: Thomson Reuters Open Calais[8] and Ontotext's Semantic Annotation[9] (Kaur and Chopra, 2016). Both of these services were looked at, they offered similar features and appear to work in similar ways (in that both use data and tags related to Wikipedia in their algorithms to tag the documents). However Open Calais had more detailed documentation and specifically offered social tags to classify the documents whereas OntoText hadn't released the document classification part of their algorithm yet. Open Calais was therefore used as the main tagging service; the implementation of which will be detailed in Section 5.5.

5.4.5.2 Ontologies

An ontology is essentially a vocabulary or dictionary for the Semantic Web (W3C, 2015). It provides a formal definition of the typical terms used within a domain, and details the hierarchy of classes that are used to represent the different levels of restrictions and relationships within a specific domain. Ontologies in Chemistry, Physics and Biology were searched for to use in this project.

Five main chemistry ontologies were identified. Three had been created by the Royal Society of Chemistry, RXNO[10] - Named Reactions Ontology, CMO[11] - Chemical Methods Ontology and MOP[12] - Molecular Processes Ontology. The other two were ChEBI[13] - Chemical Entities of Biological Interest and CHEMINF - terms commonly used in cheminformatics. The first three were all implemented in this proof of concept project, unfortunately the other two were too big to read into the system (which will be explained further in Section 5.5. Several biology ontologies were also identified and the two that were small enough to be read into the system were PO[14] - Plant Ontology and CL[15] - Cell Ontology. Physics wise the only ontology that was identified was an Astrophysics Ontology[16]. The full list of ontologies can be viewed in Appendix G.

[8] `http://www.opencalais.com/`

[9] `http://tag.ontotext.com/`

[10] `http://www.obofoundry.org/ontology/rxno.html`

[11] `http://www.obofoundry.org/ontology/chmo.html`

[12] `http://www.obofoundry.org/ontology/mop.html`

[13] `https://www.ebi.ac.uk/chebi/`

[14] `http://www.obofoundry.org/ontology/po.html`

[15] `http://obofoundry.org/ontology/cl.html`

[16] `https://www.astro.umd.edu/~eshaya/astro-onto/ontologies/physics.html`

5.4.5.3 Automatic Chemical Recognition

With respect to automatically recognising chemicals (and marking them up for metadata purposes) three main chemical recognition services were identified: OSCAR[17], GATE[18] and ChemicalTagger[19]. OSCAR and Chemical Tagger are both natural language processing tools that aim to identify and mark-up chemicals in text. ChemicalTagger actually uses OSCAR in its system so ChemicalTagger was used to get a wider range of identification. GATE is a more general natural language processing tool but it does have a ChemTagger extension so this was also used alongside ChemicalTagger in the hope that between them most of the chemicals featured in the user's text would be automatically identified.

5.4.5.4 Searching

The advanced semantic search wasn't implemented in this first iteration due to its dependency on the must have requirements, so the searching technologies used in this first prototype were quite basic with the aim to improve them based on user feedback about both how they wanted the search to work and how they wanted their documents to be tagged (with what weightings) and marked up in the first place. Therefore, the searching was initially implemented using pure term frequency (Manning et al., 2008).

For each search term t, a term frequency score is computed based on the *weight* of that term, which is based on the number of occurrences in the document d which equates to $\mathrm{tf}_{\mathrm{t,d}}$. The same operation is then applied to the title T $\mathrm{tf}_{\mathrm{t,T}}$. These two are added together for each term in the search box to give an overall term frequency score for each document, and then this score orders the results. Two slightly advanced search options are also offered: text:searchTerm and title:searchTerm should the user wish to restrict the searching to just the text or title.

5.5 Semanti-Cat

This section will detail the architecture and implementation of Semanti-Cat, followed by a walkthrough of the different areas of the web application. The link to the GitHub repository where Semanti-Cat is stored can be found in Appendix F.

[17] `https://bitbucket.org/wwmm/oscar4/wiki/Home`

[18] `https://gate.ac.uk/`

[19] `http://chemicaltagger.ch.cam.ac.uk`

5.5.1 Architecture & Implementation

After identifying the tools for this project, the first prototype was created as a Java Web Application. Java was mainly used to make use of the Apache Jena[20], an open source java framework for creating semantic web applications. The high architecture of the system is detailed below in Figure 5.2.

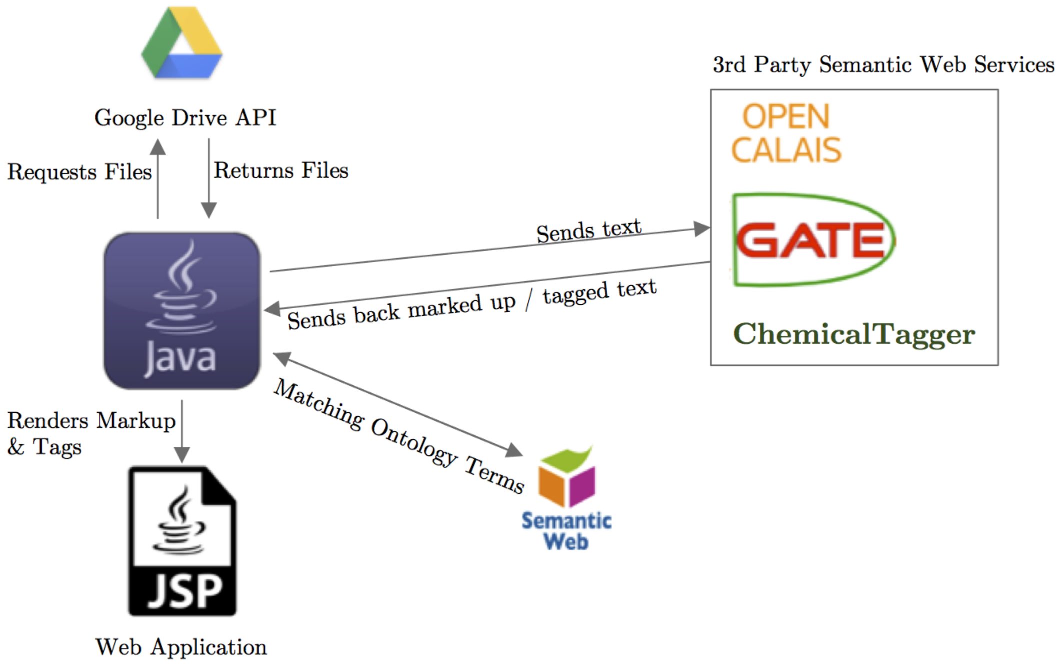


FIGURE 5.2: Architecture of the Semantic Layer

As Figure 5.2 illustrates, the web application uses the Google Drive API to obtain user selected files from their Google Drive and then runs these documents through the Open Calais API and the two chemical recognition services ChemicalTagger and GATE. Finally, these documents were also matched against the ontologies detailed in Section 5.4.5.2 to see what terms from there existed in the document.

5.5.1.1 OpenCalais API

The OpenCalais API was accessed using an HTTPClient Post to send a request containing the documents to tag, and the API sends back the document marked up with social tags that can be extracted out. These request headers had the following parameter options and types that were defined in the OpenCalais User Guide[21].

[20] https://jena.apache.org/

[21] http://www.opencalais.com/wp-content/uploads/folder/ThomsonReutersOpenCalaisAPIUserGuideR11_3.pdf

- **X-AG-Access-Token**: The parameter needs to be the license key for the API, this has been hidden and represented by stars in the above code snippet.
- **Content-Type**: This parameter needs to be the mimeType of the document(s) that are being sent, the options are text/html (web pages), text/xml (xml documents), text/raw (clean, unformatted text), and application/pdf (PDF files as binary streams, only available to premium users). Text/raw was chosen for this first implementation, as this project used the free version of the API, and based on the results of the first focus group the participants were most likely to produce word or PDF documents, which could both be converted to .txt documents.
- **omitOutputtingOriginalText**: This parameter can either be set to true or false to determine whether the original document content is sent back in the response or not, in this instance receiving the original text was not necessary as it was being sent from the same system that received it, so this was set to true.
- **x-calais-contentClass**: This parameter can be used to specify the genre of the input document, however the only options are none (default value), news (news stories) or research (for research papers, but only supported for PDF files). Given that the papers would not be news stories, and this project used the free version which did not support PDF, this wasn't set and was left to the default value.
- **x-calais-language**: This parameter defines the language of the document to be tagged and could currently be set to English, French or Spanish. Given that the evaluation of this software was to be conducted at an English University, this was set to "English".
- **x-calais-selectiveTags**: This parameter defines the metadata tag types to be used to categorise the document. There were a number of options for this parameter but the most relevant was socialtags, which attempt to classify the document as a whole, based on Wikipedia folksonomy.
- **outputFormat**: This parameter defines the output format of the response to this HTTPClient Post. The options were xml/rdf (default), application/json and text/n3. All three of these options could have been parsed to extract the tags, so the default value was chosen.

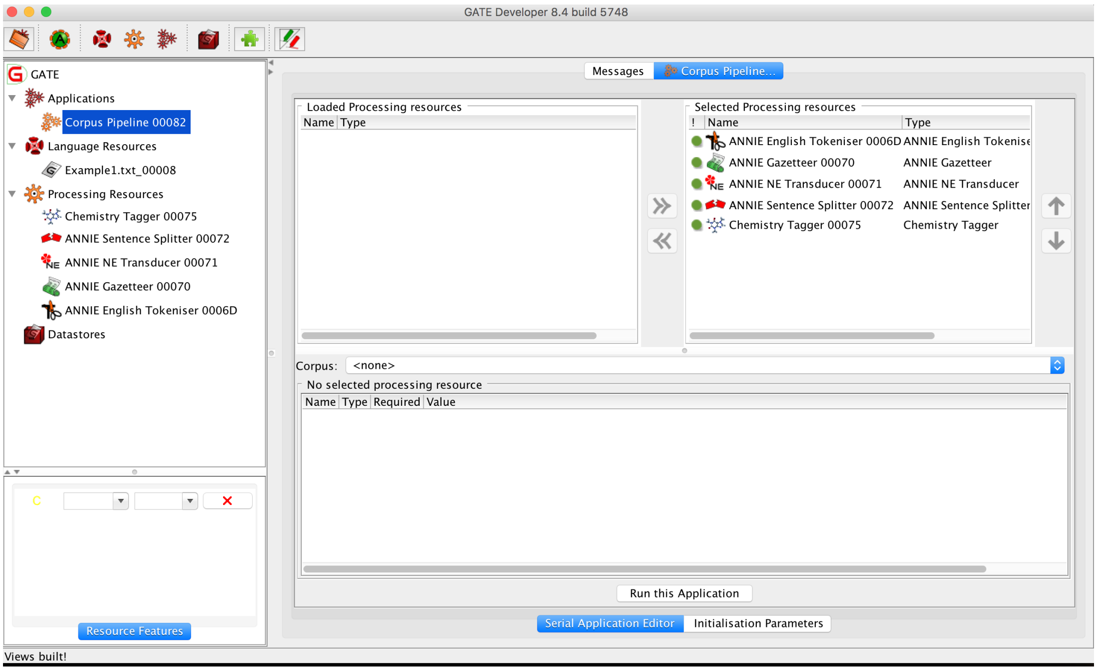
Therefore, the request headers that were set for the HTTPClient Post are detailed below in the code snippet:

```
post.setRequestHeader("X-AG-Access-Token", "********");
post.setRequestHeader("Content-Type", "text/raw");
post.setRequestHeader("omitOutputtingOrignialText", "true");
post.setRequestHeader("x-calais-language", "English");
post.setRequestHeader("x-calais-selectiveTags", "socialtags");
post.setRequestHeader("outputformat", "xml/rdf");
```

The social tags were then extracted from the HTTPResponse and assigned to the document as tags.

5.5.1.2 GATE

GATE can be used in two main ways for text processing, firstly as a standalone application (GATE Developer), and secondly within another program (GATE Embedded). To use GATE Embedded to mark-up documents, there are two main ways of achieving this using Java, the GATE environment can be fully constructed from scratch in Java, or a version of the GATE Application that denotes the appropriate setup and plugins to use can be loaded in and constructed[22]. For simplicity, the second option was chosen and the GATE Application illustrated in Figure 5.3 was loaded in.



Selected Processing resources

!	Name	Type
●	ANNIE English Tokeniser 0006D	ANNIE English Tokenise
●	ANNIE Gazetteer 00070	ANNIE Gazetteer
●	ANNIE NE Transducer 00071	ANNIE NE Transducer
●	ANNIE Sentence Splitter 00072	ANNIE Sentence Splitter
●	Chemistry Tagger 00075	Chemistry Tagger

Corpus: <none>

No selected processing resource

Name	Type	Required	Value

FIGURE 5.3: GATE Application Setup

This basic setup uses ANNIE which is GATE's information extraction system (a Nearly New Information Extraction System). The ANNIE English tokeniser is loaded in to split the document into separate numbers, punctuation and words of different types (tokens), the ANNIE gazetteer is used to identify entity names in the text based on a set of lists that has been pre-produced, the ANNIE sentence splitter is used to split the document up into sentences and the ANNIE Transducer handles split outputs of tokens such as ‘don’t’ which would traditionally be split into three tokens ‘don’ ‘” and ‘t’ to ensure that they work properly with the English tokeniser. ANNIE handles the basic natural language processing, and is required by GATE to run the application. The final processing resource

[22] `https://gate.ac.uk/sale/tao/splitch7.html#chap:api`

is Chemistry Tagger, which is a plugin built to extract and tag chemical compounds and chemical elements. GATE returns the documents with partial XML style mark-up around the identified chemicals.

5.5.1.3 ChemicalTagger

ChemicalTagger is a Natural Language Processing tool created by Hawizy et al. (2011) that takes a string of text as an input and creates an XML document. It tokenises the document using a combination of OSCAR4, domain-specific regex and English Taggers to form a tree structure that is then converted to XML tags for the different types of chemicals, experimental actions, and English phrases.

This service is simple to implement, the jar files can be downloaded from the ChemicalTagger website and run with these lines of code:

```
ChemistryPOSTagger chemPos = ChemistryPOSTagger.getDefaultInstance();
ArrayList<ELNDocument> docs;

for(int i = 0; i < docs.size(); i++){
     POSContainer pc = chemPos.runTaggers(docs.get(i).getFileContent(), true);
     ChemistrySentenceParser csp = new ChemistrySentenceParser(pc);
     csp.parseTags();
     documents.get(i).setChemTaggerDocument(csp.makeXMLDocument());
}
```

This code creates a default instance of the ChemistryTagger which calls a tokeniser, the oscarTagger (which tags chemicals), the domain specific regexTagger that looks for the formation of chemistry actions, and the openNLPTagger which uses natural language processing to tag different English phrases such as nouns and verbs. The code then runs the different taggers on the document, followed by running the sentence parser to split up the document into sentences. Finally, the output XML Document is created containing the original document text split up into XML tags. This document is then parsed using a standard SAX parser which strips out the OSCAR chemical tags and actions.

5.5.1.4 Ontologies

The ontology matching code was written for this project, and looked for matches of terms within the different ontology classes and these were also added as tags in the first instance. Where the ontology contained descriptions of the terms these were also stored so that they could be displayed when the matched ontology terms were hovered over. In the first iteration of this project all the identified tags, chemicals and ontology terms were kept and stored as metadata in relation to the documents. The logic behind this was to show the participants in the software evaluation focus groups a wide range of potential tags and chemicals to see which ones they actually would use.

For each ontology, a new object of this projects defined class type Ontology was created and the ontology was loaded in using Jena, which is simply illustrated in the code snippet below:

```
private OntModel myModel = null;
public Ontology(String fileName, String base){
     FileReader ontoReader = null;
     try{
          ontoReader = new FileReader(fileName);
     }catch(FileNotFoundException fe){
          System.out.println(fe.getStackTrace());
     }
     myModel = ModelFactory.createOntologyModel();
     myModel.read(ontoReader,base);
}
```

For each ontology, the classes and annotations were iterated over to produce a set of classes with their descriptive labels (which would later be used as the tooltip text for the ontology tags). A simple algorithm was then run over each document to match the ontology terms within the documents, iterating over each ontology term to see if it existed within the document, and if so assigning it as a tag and producing the tooltip mark-up such that the front-end web interface of Semanti-Cat would have the text to use in the tooltip. This was done in a very basic way for this first iteration to test how many ontology terms would be matched and to see which tags the participants of the evaluation focus groups thought were useful.

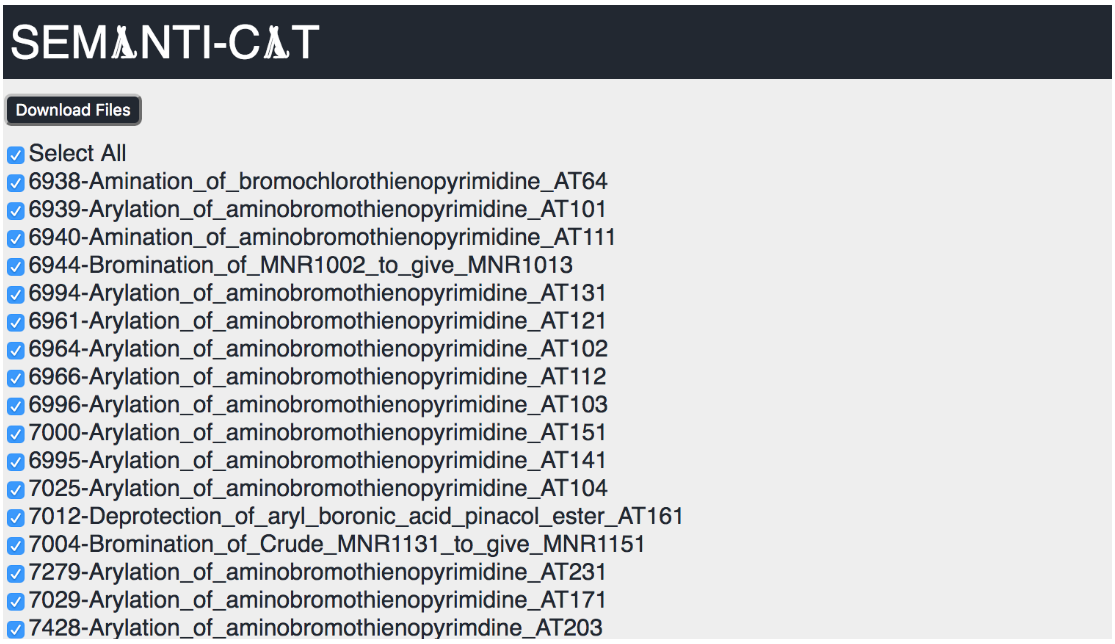
5.5.2 Semanti-Cat Walkthrough

This section shows a walkthrough of the different areas of Semanti-Cat. For the purposes of these screenshots and to initially test the system, a set of posts were taken from the Open Source e-maleria blogs[23].

5.5.2.1 Obtaining Documents

This web application used the Google Drive API to connect to a Google Drive account and list the documents that were either owned or shared with the account holder. Any or all documents can be selected to then be downloaded into the system and tagged/marked up and presented in Semanti-Cat.

[23] http://malaria.ourexperiment.org



SEMANTI-CAT

Download Files

☑ Select All
☑ 6938-Amination_of_bromochlorothienopyrimidine_AT64
☑ 6939-Arylation_of_aminobromothienopyrimidine_AT101
☑ 6940-Amination_of_aminobromothienopyrimidine_AT111
☑ 6944-Bromination_of_MNR1002_to_give_MNR1013
☑ 6994-Arylation_of_aminobromothienopyrimidine_AT131
☑ 6961-Arylation_of_aminobromothienopyrimidine_AT121
☑ 6964-Arylation_of_aminobromothienopyrimidine_AT102
☑ 6966-Arylation_of_aminobromothienopyrimidine_AT112
☑ 6996-Arylation_of_aminobromothienopyrimidine_AT103
☑ 7000-Arylation_of_aminobromothienopyrimidine_AT151
☑ 6995-Arylation_of_aminobromothienopyrimidine_AT141
☑ 7025-Arylation_of_aminobromothienopyrimidine_AT104
☑ 7012-Deprotection_of_aryl_boronic_acid_pinacol_ester_AT161
☑ 7004-Bromination_of_Crude_MNR1131_to_give_MNR1151
☑ 7279-Arylation_of_aminobromothienopyrimidine_AT231
☑ 7029-Arylation_of_aminobromothienopyrimidine_AT171
☑ 7428-Arylation_of_aminobromothienopyrimdine_AT203

FIGURE 5.4: Obtaining Documents

Once this step is complete, the user will be directed to the main page of the system where they can view the documents with mark-up and tags.

5.5.2.2 Viewing Documents

This backend view illustrates how all of the documents have been marked up and tagged and what chemicals have been identified. The different tabs down the left hand side show the following pieces of information for each document:

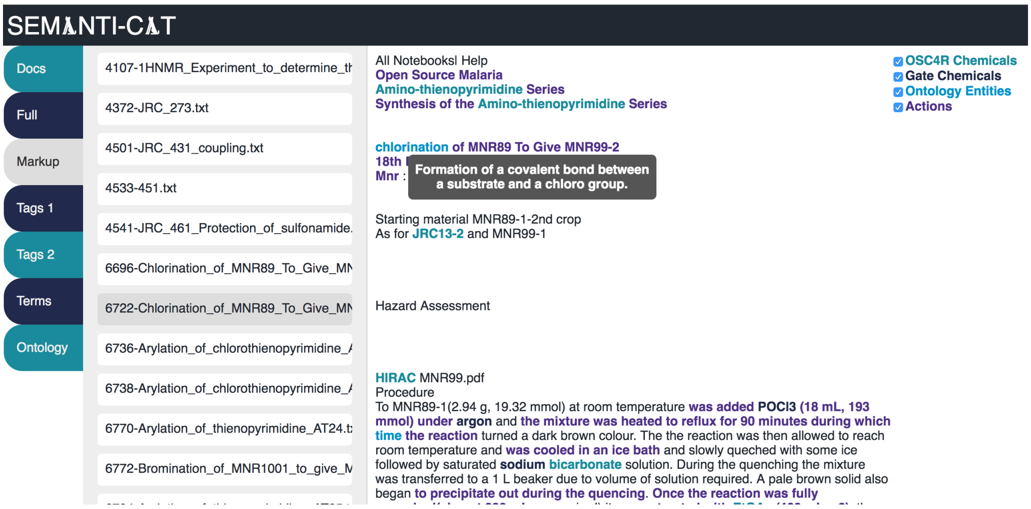
- **Docs** - This tab shows the document before it had any edits made to it.
- **Full** - This tab shows the document with both the mark-up applied to it and the tags
- **Markup** - This tab shows the document mark-up (which chemicals and actions have been identified and which ontology terms have been matched, and hovers to show actions, where things have been identified as chemicals and ontology definitions).
- **Tags 1** - This tab shows the different sets of tags in their categories.
- **Tags 2** - This tab shows the different sets of tags split between OpenCalais which are topic based tags, and Gate/ChemicalTagger/Ontology which have the chemicals and terms.
- **Terms** - This tab shows the term frequency of the terms identified in the different ontologies for that document.

- **Ontology** - This tab shows chord diagrams showing the co-occurrence of terms in the different ontologies for all the documents.

The following sections contain screenshots and descriptions of some of the tabs with some of the example documents:

5.5.2.3 Markup Tab

Figure 5.5 illustrates the mark-up tab. This tab shows the output of the document marked up with the different types of tags. This allows the participants to see which tags have been pulled out by which service (as the colours are defined in the top right-hand corner) and also the actions that have been identified by ChemicalTagger. On this page, the user can turn the mark-up on and off for the different services, for two reasons. Firstly, it allows them to turn them off if they do not wish to see tags from a particular service, but secondly it allows them to see where there might be overlaps or where certain terms have been tagged by more than one tagging service. This mark-up tab also facilitates tooltips, this figure shows the tooltip that appears when the user hovers over the word chlorination, with the text of the tooltip coming from the descriptive label assigned to the class chlorination in the Molecular Processes Ontology. These descriptions are to provide further information about the tags where possible.


SEMANTI-CAT
Docs
Full
Markup
Tags 1
Tags 2
Terms
Ontology
4107-1HNMR_Experiment_to_determine_th
4372-JRC_273.txt
4501-JRC_431_coupling.txt
4533-451.txt
4541-JRC_461_Protection_of_sulfonamide.
6696-Chlorination_of_MNR89_To_Give_MN
6722-Chlorination_of_MNR89_To_Give_MN
6736-Arylation_of_chlorothienopyrimidine_A
6738-Arylation_of_chlorothienopyrimidine_A
6770-Arylation_of_thienopyrimidine_AT24.t
6772-Bromination_of_MNR1001_to_give_M
☑ OSC4R Chemicals
☑ Gate Chemicals
☑ Ontology Entities
☑ Actions

All Notebooksl Help
Open Source Malaria
Amino-thienopyrimidine Series
Synthesis of the Amino-thienopyrimidine Series

chlorination of MNR89 To Give MNR99-2
18th
Mnr :

Formation of a covalent bond between a substrate and a chloro group.

Starting material MNR89-1-2nd crop
As for JRC13-2 and MNR99-1

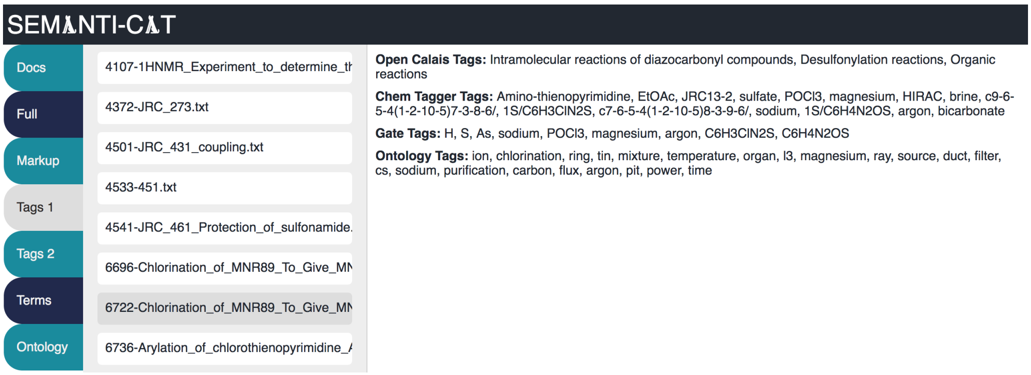
Hazard Assessment

HIRAC MNR99.pdf
Procedure
To MNR89-1(2.94 g, 19.32 mmol) at room temperature was added POCl3 (18 mL, 193 mmol) under argon and the mixture was heated to reflux for 90 minutes during which time the reaction turned a dark brown colour. The the reaction was then allowed to reach room temperature and was cooled in an ice bath and slowly quechled with some ice followed by saturated sodium bicarbonate solution. During the quenching the mixture was transferred to a 1 L beaker due to volume of solution required. A pale brown solid also began to precipitate out during the quencing. Once the reaction was fully

Figure 5.5: Semanti-Cat - Markup tab, illustrating highlighting showing where the different libraries have identified tags and ontology terms, and tooltips with ontology explanations

5.5.2.4 Tags Tab

Figure 5.6 illustrates the tags 1 tab.


SEMANTI-CAT
Docs
Full
Markup
Tags 1
Tags 2
Terms
Ontology
4107-1HNMR_Experiment_to_determine_th
4372-JRC_273.txt
4501-JRC_431_coupling.txt
4533-451.txt
4541-JRC_461_Protection_of_sulfonamide.
6696-Chlorination_of_MNR89_To_Give_MN
6722-Chlorination_of_MNR89_To_Give_MN
6736-Arylation_of_chlorothienopyrimidine_/

Open Calais Tags: Intramolecular reactions of diazocarbonyl compounds, Desulfonylation reactions, Organic reactions

Chem Tagger Tags: Amino-thienopyrimidine, EtOAc, JRC13-2, sulfate, POCl3, magnesium, HIRAC, brine, c9-6-5-4(1-2-10-5)7-3-8-6/, 1S/C6H3ClN2S, c7-6-5-4(1-2-10-5)8-3-9-6/, sodium, 1S/C6H4N2OS, argon, bicarbonate

Gate Tags: H, S, As, sodium, POCl3, magnesium, argon, C6H3ClN2S, C6H4N2OS

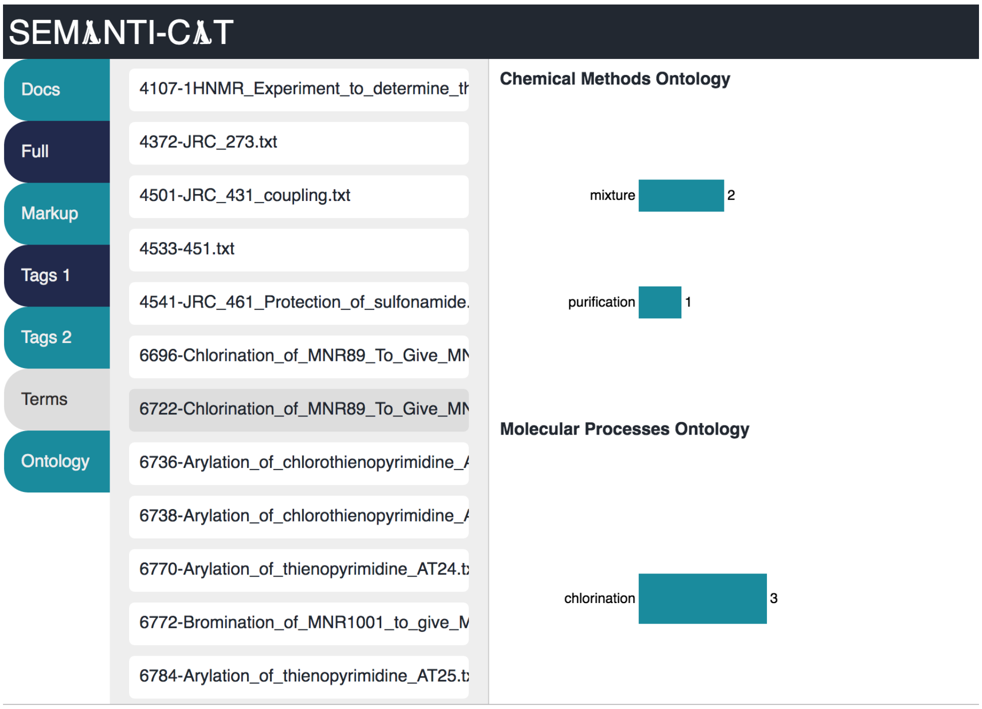
Ontology Tags: ion, chlorination, ring, tin, mixture, temperature, organ, I3, magnesium, ray, source, duct, filter, cs, sodium, purification, carbon, flux, argon, pit, power, time

FIGURE 5.6: Semanti-Cat - Tags 1 tab, shows the tags / chemicals from the different services

This tab displays lists of the different tags split up by tagging service (OpenCalais, ChemicalTagger, GATE and the Ontology Tags). This allows participants to see a quick view of all of the different tags assigned to their documents such that they can understand how it has currently been categorised and comment on how well this has been done. The tags 2 tab is very similar it just splits the tags into two groups OpenCalais tags, and the rest of the tags. This is because the OpenCalais tags are terms that won't necessarily be in the document itself as they categorise the document, whereas GATE and ChemicalTagger recognise chemicals that exist in the documents and use them as tags and similarly the ontology tags are also formed of terms that exist in the document. These tabs are meant to provide the participants who evaluate the software a quick and easy way of understanding how their documents have been tagged.

5.5.2.5 Terms Tab

Figure 5.7 illustrates the terms tab.



4107-1HNMR_Experiment_to_determine_th
4372-JRC_273.txt
4501-JRC_431_coupling.txt
4533-451.txt
4541-JRC_461_Protection_of_sulfonamide.
6696-Chlorination_of_MNR89_To_Give_MN
6722-Chlorination_of_MNR89_To_Give_MN
6736-Arylation_of_chlorothienopyrimidine_A
6738-Arylation_of_chlorothienopyrimidine_A
6770-Arylation_of_thienopyrimidine_AT24.tx
6772-Bromination_of_MNR1001_to_give_M
6784-Arylation_of_thienopyrimidine_AT25.tx

Figure 5.7: Semanti-Cat - Terms tab, shows which terms have been identified from the different ontologies

This shows the users which terms from their document have been pulled out from the different ontologies. The participants of the software evaluation study are informed of what an Ontology is and this tab is here to show them which types of ontologies have identified terms in their document such that they can hopefully gain a better understanding of why these terms have been pulled out. In addition, it will be interesting to see for example how many cross discipline terms will be pulled out, e.g. a chemical term in a biology based document, or a physics term in a chemical document.

5.5.2.6 Ontology Tab

Figure 5.8 illustrates the ontology tab.

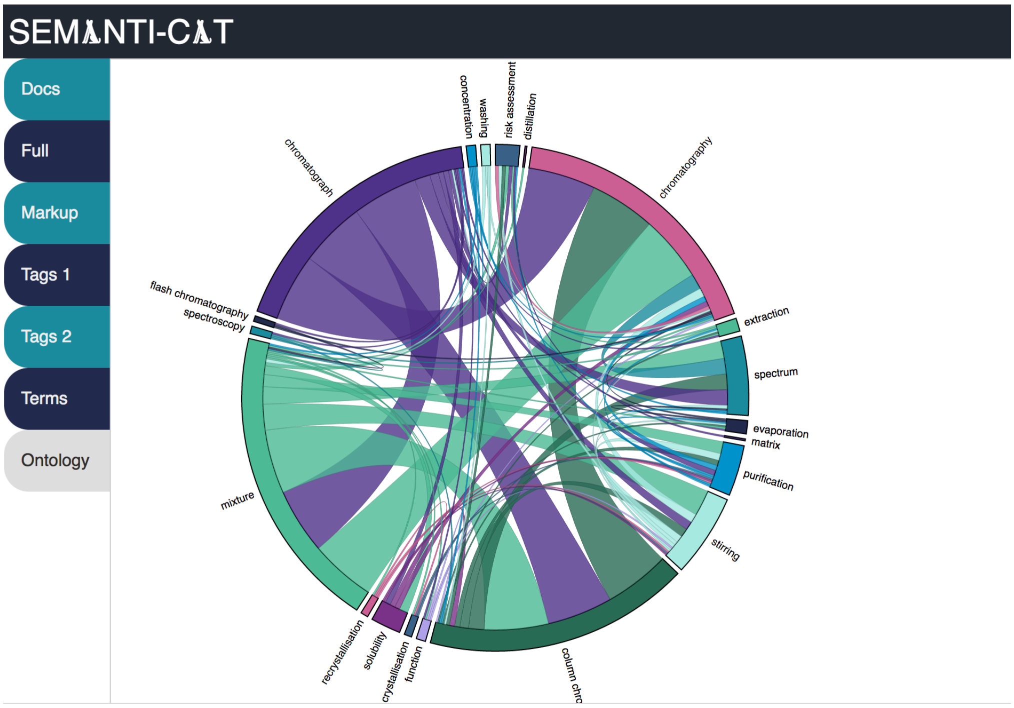


FIGURE 5.8: Semanti-Cat - Ontology Tab, shows the term co-occurrence across all the documents in the different ontologies

The main focus of Semanti-Cat is to provide the document annotations; however this tab was added in as part of this prototype to show users' which terms commonly occur across their set documents in Semanti-Cat. This will enable the users to see the tags assigned to their documents in relation to the tags assigned to others documents. Developing ontologies and tagging mechanisms to effectively tag documents is an area that remains unsolved, and understanding the co-occurrence of the terms across a set of documents (particularly across a specific domain) will aid with some of the could have requirements, to build up a set of linked terms to start forming predictions for related documents.

5.6 Conclusions

Semanti-Cat will be evaluated in focus groups consisting of physicists, chemists and biologists to assess and discuss the different aspects of its functionality. This has been deliberately coded in a generic way to allow the participants to discuss the way the documents have been tagged and marked up, to facilitate discussions of what they want from these pieces of functionality. The feedback of these evaluations will form the new implementation of the next iteration of Semanti-Cat which will be detailed as part of the future work in Chapter 7.

Chapter 6

Software Evaluation User Studies

This section will detail the software evaluation focus groups that took place to both evaluate the proof of concept software and to discuss at what point in the lab process, scientists require additional software support.

As discussed in Chapter 2, whilst many user studies have been conducted around ELNs; there aren't many studies that perform user research of this magnitude, both before and after developing a proof of concept system. One of the notable studies to undertake this was the Dial-a-Molecule iLabber pilot project detailed in (Kanza et al., 2017), (the data analysis of which was performed as part of this project and detailed in Chapter 4). This flow of user study is very important as it enables researchers to capture a wider range of information as they can get user feedback more continuously and enhance their understanding of the entire process of developing user software. In this instance, previous user studies from this thesis and all of the studies detailed in (Kanza et al., 2017) have been collated together to produce the proposed proof of concept system that was detailed in Chapter 5. This was achieved by taking all the features listed as desired by the study participants from the initial user studies and the Dial-a-Molecule studies and mapping them to the original Figure 2.2 from Chapter 2 to assign them to one of the three layers of the proposed proof-of-concept ELN (Notebook Layer, Domain Layer, Semantic Layer). These features were then broken down into non functional requirements, and the functional requirements for each of the three layers in Chapter 5. The requirements of the semantic layer which was created as the prototype Semanti-Cat were prioritised using MOSCOW (Van Vliet et al., 1993), and the following must have requirements were implemented:

FRS1: The system must tag/classify documents
FRS2: The system must store metadata about the documents
FRS3: The system must link to ontologies
FRS4: The system must be able to automatically identify chemicals

Whilst previous researchers such as Borkum et al. (2010) and Hughes et al. (2004b) have worked on designing ontologies to describe the different entities of an experiment, and research conducted by (Talbott et al., 2005) has looked at using Semantic Web technologies in an ELN, there are limited user studies that actually evaluate such an implementation. Drăgan et al. (2011) conducted a study of SemNotes (a semantic note taking platform that tries to link user records) to see how users found it, but this was a very generic piece of software rather than one that was aligned with a specific discipline, and didn't have the same type of features as the ones listed above. This study therefore looked to evaluate how these requirements had been implemented in the proof of concept study, with five main aims:

- To investigate what scientists actually wanted from a semantic layer of an ELN both in terms of features and integrated layout.
- To investigate what scientists thought of the overall concept of the layered notebook approach of this software.
- To understand how the scientists needs differed across different scientific disciplines.
- To understand at which point in the lab process this type of software should be aimed at.
- To gain further understanding about why scientists are reluctant to use ELNs.

These aims directly correspond to the research questions that these focus groups were designed to help answer which are: RQ, RQ.2, RQ.3 and RQ.4. The detailed methodological approach that was outlined in Chapter 3 will be fully described, followed by the results of this study and the subsequent discussion of these results both in relation to this study and to the previous studies undertaken as part of this project. In the results section, like in Chapter 4 the participants of the studies will be referred to either as PhD students or participants, but when referring to a specific set of them by discipline then that disciplinary term will be used (e.g chemists, physicists, and biologists).

6.1 Methodology & Study Design

The software evaluation focus groups were designed in a similar way to the initial focus groups detailed in Chapter 4, to use a semi structured set of questions to encourage discussion around certain topics, but still leave the forum of discussion open enough for the participants to debate and talk around the subjects raised (Britten, 2007). The participants who took part in these focus groups are Postgraduate students from the University of Southampton, studying Physics, Chemistry or Biology. A trial focus group was run first to test the format of the focus group and see whether it made sense with regards to evaluating the software, and also to test the questions to see if they elicited the answers and subsequent discussion that was needed to answer the research questions. The participants that took part in these focus groups were a mixture of previous participants who had been involved with the initial focus groups and or lab observations, and some new participants. There were some limitations in availability of participants and timing and the overall makeup of the fifteen focus groups participants are listed below; links to the ethics application and transcriptions of these focus groups are all listed in Appendix H.

- Trial Focus Group: 3 Participants A, B & R
- Chemistry Focus Group: 6 Participants L, J, S, AP, AJ & AK
- Physics Focus Group: 4 Participants A, B, AL & AM
- Biology Focus Group: 5 Participants Q, R, AN, AO & AQ

The initial set of questions the trial participants were asked were split into three parts:

1. Study Format
2. Software Evaluation
3. ELN Behaviour

6.1.1 Study Format

The study format section was designed to introduce the study and explain how the focus group would work to see if it made sense and to ask the trial participants about how the software evaluation should be conducted, they were asked whether they thought it would be better to view the software as a group on a projector and discuss it, or if they thought it would be preferable to look at it individually on laptops. This was the layout of the study format section, which was not refined after this focus group as it was only discussed to improve how the main focus groups were organised.

1. Explain purpose of the study and software and show participants the software.
2. Ask the participants if this make sense and if they understand what they are being asked to do?
3. Ask the participants if they think it would be better to look at software on projector screen together or individually on their own laptops?

6.1.2 Software Evaluation

The software evaluation section was designed to discuss different aspects of the system. It starts by trying to understand what documents might be used in the system and to further understand the makeup of those documents. Participants are asked to send a document to be marked up in the system ahead of time, and this was partially to see what type of document they brought, and also to see what variation of documents the ELN software would need to handle. The participants were asked to describe the document they had brought and asked to detail how typical it was of their usual work. They were then asked about what proportion of their documents were made up of words/diagrams/picture. This was important to ascertain because typically images and diagrams are much harder to tag and mark-up (and create using software rather than paper) and so would affect a user's decision to create or place this type of work in an ELN system. Finally, participants were also asked what formats they typically created their documents in as the current system facilitated storing any type of document (as Google Drive allows uploading of all document formats) but with regards to tagging, currently only .txt .doc and .pdf files were supported by converting everything to .txt (although .pdf didn't always accurately extract all the necessary text to mark-up compared to .doc), and this ELN platform would need to support the different types of documents that users would want to tag and mark-up.

Following this the software evaluation section discussed the different tagging systems used in the software. The trial focus group participants were shown test documents and the tags that had been assigned to them (in the main focus groups participants would be shown the tags assigned to the documents that they had sent in), and the different graph tabs of the system to show which tags had been identified from which services and how the corpus of documents had been tagged as a whole. They were asked to identify which tagging systems were the most appropriate and inappropriate for their work, and whether they felt there were any obvious missing tags, both to understand how well they worked, and also whether different tagging systems worked better for documents of different scientific disciplines. The participants were also asked how they would personally tag their documents and whether they would make use of the options to add and remove their own tags. This was to get a better idea of how different participants would expect their documents to be tagged and how much involvement they would want to have with that process.

The topic of search was next, which aimed to explore how the participants expected a search to work, and what types of advanced search options (if any) they desired. One of the main aims of this focus group was to understand how to create an advanced semantic search to meet the user requirements from the previous studies, with the look to refine the tagging process based on user feedback and then design a search feature that would fit with the users' needs.

After search, the mark-up was discussed; Semanti-Cat had marked up the documents to show which terms and actions were pulled out by which tagging service, both to illustrate the difference between the third party services but also to show how much of the document information was extracted out into tags. Additionally, descriptions of the types of tags were put in tooltips that could be viewed by hovering over the tags. The ontology tags detailed the descriptions given in the ontology as part of their tooltip to provide the user with further information, the chemical elements that were identified had tooltips that identified them as chemical elements and the different actions pulled out by ChemicalTagger were also marked up and given tooltips that detailed what type of action they were. The participants were asked how (if at all) they would want to see their documents marked up, and whether they thought the tooltips were useful.

Finally, the users were asked what parts of the functionality demonstrated in SemantiCat would they want to see in a notebook platform, and how they would expect it to look. This was asked to facilitate discussion about using these features in an existing notebook platform and to ascertain what the users' expectations were of how tags and mark-up would be handled in that type of environment.

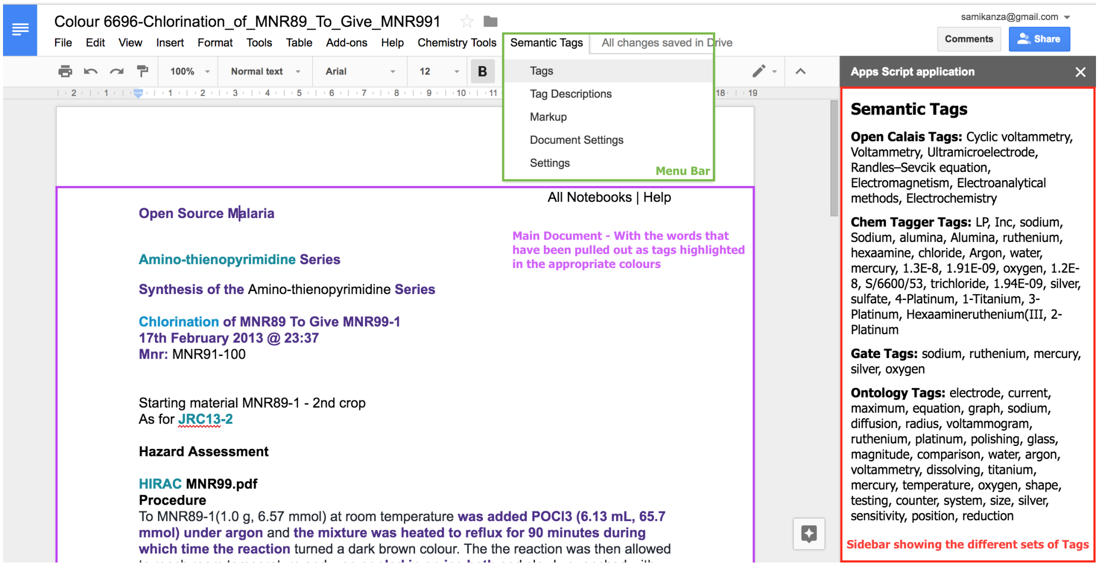
6.1.3 ELN Behaviour

This section was designed to understand whether these pieces of functionality detailed by scientists from the previous user studies (both from Dial-a-Molecule survey, and the focus groups with PhD students from the University of Southampton) would have any impact on persuading them to use this type of software, and to understand what stage in the lab process this software might belong. The participants were asked whether they agreed with one of the main outcomes of the previous studies, in that scientists have a greater tendency to use software for writing up formal papers, reports, and thesis's rather than during their lab experiments. They were then asked if an ELN aimed at that stage of the lab process would be more attractive for them to use, and whether semantic tagging and searching would impact their decision to use an ELN or whether it would encourage them to further digitise their work. They were also asked whether they thought automatic tagging of their documents would have any impact on the efficiency of their work.

6.2 Refined Methodology & Study Design

The outcomes of the trial focus group helped refine both the organisational structure of the main focus groups, and the questions. With respect to the study format section, the participants all unanimously agreed that they would rather look at the software as a group on a projector rather than looking at it individually on their own laptops. In the software evaluation section, asking about the different tags as a whole didn't facilitate the desired discussion about whether ChemicalTagger and Gate had done a decent job of auto recognising the chemicals, so a question was added in about this. Additionally, the idea of manually editing the tag descriptions as well as manually adding/removing tags was brought up in discussion in the trial focus group so a question about this was also added in. With regards to discussing the system design, the participants also stated that they found it difficult initially to understand that they were being asked to evaluate the

Semanti-Cat system with respect to its functionality, but also to conceptualise what they would expect to see in the front-end with respect to Google Docs. In order to address this, some mock-ups were produced for the front-end using Google App Scripts to make things clearer in the following focus groups. The initial mock-ups were designed using the feedback from the first focus group with the aim to iteratively build upon those designs using the main focus group feedback to form part of the future work. Figures 6.1 and 6.2 show how the front-end has been implemented with Google Docs.



Open Source Malaria

Amino-thienopyrimidine Series

Synthesis of the Amino-thienopyrimidine Series

Chlorination of MNR89 To Give MNR99-1
17th February 2013 @ 23:37
Mnr: MNR91-100

Starting material MNR89-1 - 2nd crop
As for JRC13-2

Hazard Assessment

HIRAC **MNR99.pdf**
Procedure
To MNR89-1(1.0 g, 6.57 mmol) at room temperature was added POCl3 (6.13 mL, 65.7 mmol) under argon and the mixture was heated to reflux for 90 minutes during which time the reaction turned a dark brown colour. The the reaction was then allowed

Semantic Tags

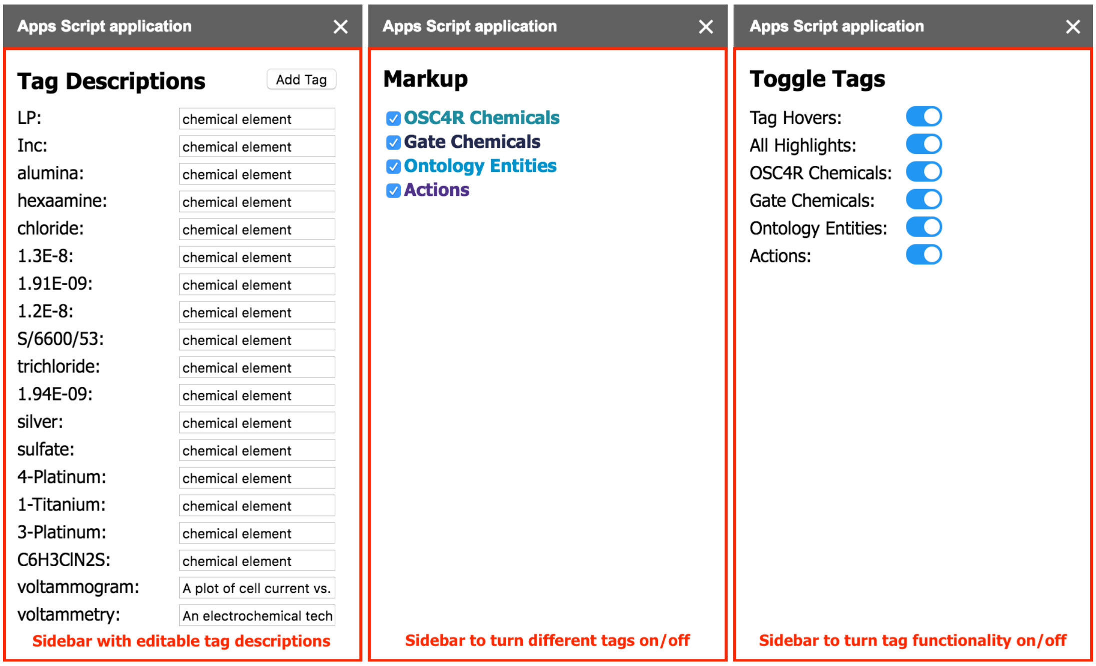
Open Calais Tags: Cyclic voltammetry, Voltammetry, Ultramicroelectrode, Randles–Sevcik equation, Electromagnetism, Electroanalytical methods, Electrochemistry

Chem Tagger Tags: LP, Inc, sodium, Sodium, alumina, Alumina, ruthenium, hexaamine, chloride, Argon, water, mercury, 1.3E-8, 1.91E-09, oxygen, 1.2E-8, S/6600/53, trichloride, 1.94E-09, silver, sulfate, 4-Platinum, 1-Titanium, 3-Platinum, Hexaamineruthenium(III, 2-Platinum

Gate Tags: sodium, ruthenium, mercury, silver, oxygen

Ontology Tags: electrode, current, maximum, equation, graph, sodium, diffusion, radius, voltammogram, ruthenium, platinum, polishing, glass, magnitude, comparison, water, argon, voltammetry, dissolving, titanium, mercury, temperature, oxygen, shape, testing, counter, system, size, silver, sensitivity, position, reduction

FIGURE 6.1: Google Front-End Mock-ups (Main Page) designed with Google App Scripts


Apps Script application
Apps Script application
Apps Script application
Add Tag
Sidebar with editable tag descriptions
Sidebar to turn different tags on/off
Sidebar to turn tag functionality on/off

Tag Descriptions

LP:	chemical element
Inc:	chemical element
alumina:	chemical element
hexaamine:	chemical element
chloride:	chemical element
1.3E-8:	chemical element
1.91E-09:	chemical element
1.2E-8:	chemical element
S/6600/53:	chemical element
trichloride:	chemical element
1.94E-09:	chemical element
silver:	chemical element
sulfate:	chemical element
4-Platinum:	chemical element
1-Titanium:	chemical element
3-Platinum:	chemical element
C6H3ClN2S:	chemical element
voltammogram:	A plot of cell current vs.
voltammetry:	An electrochemical tech

Markup

☑ OSC4R Chemicals
☑ Gate Chemicals
☑ Ontology Entities
☑ Actions

Toggle Tags

Tag Hovers:	●
All Highlights:	●
OSC4R Chemicals:	●
Gate Chemicals:	●
Ontology Entities:	●
Actions:	●

FIGURE 6.2: Google Front-End Mock-ups (Sidebars) designed with Google App Scripts

Figure 6.1 shows the main document with the custom menu bar (Semantic Tags) that links to the different pieces of functionality that map to the tags and mark-up tabs in Semanti-Cat, with the tag sidebar open. Figure 6.2 shows the other sidebars available: 'Tag Descriptions' for the users to see and edit the different descriptions for the tags; 'Markup' to turn off specific highlighted tag types; and 'Document Settings' to toggle hovers, all highlights, actions and the assignment of specific types of tags. Additionally, an overall settings option exists in the menu bar to toggle the different pieces of functionality exhibited in the document settings pane, on and off for all documents.

Based on the feedback from the trial focus group, the different pieces of functionality were put in a sidebar, with different sidebars for the tags, tag descriptions, mark-up, document settings and then an overall settings panel was also added that would change the settings for the whole notebook. These sidebars are accessible from a toolbar as demonstrated by Figure 6.1. The last question about the front-end design was slightly changed from: "What would you expect to see in the actual document if you were using a system that implemented this functionality" to "What do you think of the current design of the system? - Whilst showing the front-end mock-ups to the participants to discuss". This was changed based on creating the mock-ups to give participants a tangible design to discuss and refine, in addition to generating a discussion about the overall concept of adding this functionality to a pre-existing notebook system rather than just showing them the backend and asking them what they would expect to see integrated into their notebook software. The last set of questions detailing ELN behaviour was kept the same as this generated an informative discussion in the trial focus group. Taking into account these changes the main focus groups were redesigned to ensure that the explanations given made it clearer to the participants how they were being asked to evaluate the software and the difference between Semanti-Cat and the Google Drive front-end. The newly redesigned outline of the study will be detailed below, and the verbatim descriptions and full questions can be found in Appendix H.

6.2.1 Study Format

The motivations of the study were explained, giving the history of what has been done thus far for this PhD including explaining the initial user studies and market / literature research. The participants were informed of the interim conclusions that had been derived, that the studies had shown that there were many good reasons to digitise one's lab data, and that there was already a significant use of software among scientists, but that barriers still existed to using electronic lab notebooks. It was explained that this software had not been created to replace paper, but that it was aiming towards improving the current software offerings that were available, and was designed to aid in the writing up stage of the lab process. Following that the motivations and decisions behind building an ELN upon an existing cloud platform was explained, and the three-layer proposal was shown to the participants and explained to help them understand not only the overall vision of this software, but what pieces of functionality they would be evaluating. The Semanti-Cat backend proof of concept system and the front-end mock-ups in Google Drive were explained, alongside the overall aims and format of this focus group so that the participants would be fully informed before answering the questions.

6.2.2 Software Evaluation

The software evaluation stage of the study was broken into five different areas: general system, tags, search, mark-up, and design of system.

For the general system, the users were asked about what documents they had brought and how typical a representation of their working documents it was. They were also asked about the general makeup of their documents between words pictures and diagrams, and what formats they were typically created in.

Following these questions, the different tagging services and automatic chemical recognition services were explained before the users were asked which tagging services they felt were most and least useful, and if any obvious tags were missing. They were also asked how they felt ChemicalTagger and GATE fared with regards to automatically recognising their chemicals, and then they were asked how they would have tagged the document if they had done it themselves and if they would make use of an option to add and remove the tags themselves. The chemical recognition question was added in to ensure that the participants understood that these were automatically recognising chemicals and pulling them out as tags as opposed to the other two tagging sets, which were aiming to pull out related terms.

The participants were then asked about their searching preferences. They were asked about their assumptions towards how a search feature would work, and how they would expect the results to be prioritised, alongside questions about advanced search functionality.

Following this the participants were asked about the mark-up tab, and whether they could see any merit in highlighting the tags that the system had pulled out, and if they thought seeing the associated descriptions of some of the terms as hover tooltips would prove to be useful. They were also asked if they would make use of the ability to edit the tags and chemical descriptions themselves, and whether they had any better suggestions for how to mark-up the documents.

Finally, in this set of questions the participants were asked about their opinions towards the design of the system. They were shown each different pane in Google Drive that could be opened using the custom menu, and explained how they showed the functionality implemented in Semanti-Cat, and asked about their opinions on these initial mock-ups.

6.2.3 ELN Behaviour

After discussing the software, the participants were then asked some questions about ELNs in general and what impact they thought this software could have. They were asked about whether they agreed with the previous findings of this project that scientists have a greater tendency to use software for writing up formal papers, reports and thesis rather than during their lab experiments to take notes. They were then asked if they would be more inclined to consider using an ELN if it was aimed at this stage of the lab, and or if it semantically tagged their work and provided a more enhanced search. They were also asked if these features would encourage them to further digitise their work, and what impact they thought it would have on the efficiency of their work. Asking the participants if they wished to make any further comments finished these discussions.

6.3 Results

This section will detail the results of the focus groups. They have been broken down into the different areas of questions that were asked in the focus groups.

6.3.1 General System

Prior to taking part in these focus groups, the participants were asked to send a piece of their work to be marked up and tagged in the proof of concept system. When the system was developed some chemistry blogs were used as test data but in order to see how well it actually met their requirements, it needed to be trialled with their own work. The information the participants were given in the participant information sheet with regards to sending a document was: *"You will be asked to send a scientific document (either one of your own or one that you feel bears similarity to the kind of work you do."* They were deliberately given limited guidance to allow them freedom to send over a document that they felt would be appropriate for an ELN. Both results from previous recorded studies in literature (Shankar, 2007) and the initial user studies from this project detailed in Chapter 4 illustrated that the notes that the PhD students produced were very personal and differed greatly. It is important to fully trial this system with a range of documents to ensure that the contrasting needs of different scientists can be met.

The documents sent over ranged from sections of participant's theses, a research paper, electronic supplementary information for a research paper, lab reports, experiment write ups, a blog post, and a literature review. They also varied in length and format; some were Word documents, some were PDFs, and some were pictures of handwritten pages from their lab books that had to be converted into an electronic format to be run through the system. Similarly to the results of the first set of focus groups detailed in Chapter

4, there was also a variation in the contents of the documents. Some were pure text but contained material the participants noted that wouldn't necessarily want tagged such as reference lists and some contained figures, graphs, tables or pictures. In the original focus groups both biologists and physicists showed some uniformity in their work and the chemists were more diverse; in this set of groups, the biologists also showed uniformity in the documents they sent over. The biologists' documents consisted of two experimental write ups, one literature review and two thesis sections, and all were predominantly words with a few tables, graphs and figures. The physicists however showed less uniformity in the work they sent over, that ranged from a large research paper, two handwritten lab book pages and a sample report. These differed in makeup, as the research paper was about 60% words to 40% figures and equations, the handwritten lab pages were both different to each other, one was very short with a few words and a sketched diagram, and the other was longer and entirely words. The sample report had a roughly equal level of words and figures. the chemists showed a similar level of diversity to the original studies, their documents comprised of an ESI (Electronic Supporting Information), two thesis sections, a blog post, a handwritten lab book page and a full experiment write-up. The ESI was about 60% graphs and figures to 40% words, the thesis sections were almost entirely words with one figure, the experiment write-up was predominantly figures and graphs with some words; the blog post was mostly structure diagrams and a few words and the handwritten lab book page was quite short with a few words and some noted down filenames.

These documents themselves already illustrate not only the range of pieces of work that PhD students produce but also the different content that an ELN platform would need to handle, including content that is both challenging to tag and mark-up or even digitise in the first place in some instances (figures, sketches or diagrams) and content that users may not want tagged such as reference lists. Further to this, asking the participants about what formats they typically created their documents in and what their general mark-up was with regards to words and pictures and diagrams elicited a similar diversity. The biologists were again fairly uniform, agreeing that they predominantly produced documents that were mostly words with some figures and tables, and that their figure captions would be quite detailed. They all said that they created their documents in word, and used software such as PRISM, GraphPad or Excel to model their data and produce their graphs. This type of work could fit quite well with the overall proof of concept ELN as Google Drive facilitates Word and Excel like programs, and using it wouldn't change their current habits. Additionally, documents in .doc format and documents that are predominantly words with detailed captioned figures are easier to tag as most of the relevant information can be easily extracted from the document.

6.3.2 Tagging

The next item of discussion in the focus groups were the tagging and auto chemical recognition systems that were used in the system (these were described to the participants in the focus group questions / study design detailed in Section 6.2.2).

There were some variances among the different disciplines with regards to how the participants felt about the different services but there were also some common themes that came out of these discussions. Almost every participant regarded Open Calais as too generic and broad. In each focus group one person said that out of all four services Open Calais was probably the most useful, but even they prefixed that statement with an opinion that they were still too broad. One chemistry participant (Participant J) commented "if it was for other people looking into a much larger system than this then it could be more useful" identifying that the generic descriptions the group had been given by Open Calais of 'chemistry', 'organic chemistry' and 'electro chemistry' weren't very useful when that was their scientific specialisation and in all likelihood all of their documents would be tagged as such, whereas in a larger system of varying disciplines or sub disciplines these tags would provide a more useful differentiation. These comments show how important the performance expectancy element of UTAUT is, as the participants were actively looking for services that would add value to their work. Additionally, Open Calais could be used to identify the core disciplines to narrow down the types of tags attributed to the documents.

Similarly, when evaluating the ontology tags, most participants said that they found them too generic and also that too many tags were pulled out from them. For example, Participant AP's four-page document, Semanti-Cat pulled out 32 tags which were deemed too many to helpfully classify a document. Additionally, in this document 'voltammetry' was identified as an ontology term, whereas Open Calais identified 'cyclic voltammetry' as a tag, which Participant AP said, was more helpful. In Participant S's 12-page document, 75 ontology tags were extracted and they highlighted six terms that they felt would be tags that they would use, that were either specific or highly related. A full list of the participants documents, with details about the documents lengths, and the number of tags identified from each service can be found in Appendix I.

Some participants thought that the ontology tags were useful in that they picked out the broad themes, although as Participant S stated "I'm not going to search by theme, I know what I do and I want to know which experiment it is". Ontology tags could also be used in a similar fashion to identify terms to assign a general discipline to the document, which could then call different tagging services or put certain restrictions on tagging services to only pull out the relevant tags. Six different ontologies were used in the proof of concept system (3 Chemistry, 2 Biology and 1 Physics), and the nature of the ontologies hierarchical structure means that some of the top level terms will be very generic words like 'ring' and graph' for example from the physics ontology.

Additionally, there were some comments that not all of the ontology tags were relevant. Some of the more generic tags assigned to documents of one discipline were from ontologies not specific to that discipline, but there are some cross overs in scientific terms especially at the top level of the hierarchy so these were still being identified, such as 'group' in the reactions ontology. Furthermore, there are some ontology terms that have more than one meaning, for example 'current' is a term in the physics ontology, and mixture a word in the chemical methods ontology; these words also have other meanings whereby it would not merit them being highlighted as a tag. This illustrates that a further level of natural language processing is necessary to both narrow down appropriate terms and rule out ones that are meant in a different context. It could make for a better tagging system if the Open Calais tags were used to narrow down which ontologies should be considered and then write some extra natural language processing methods to only tag certain terms.

With regards to chemical recognition tools, ChemicalTagger and GATE, the participants were unanimous in that they didn't find GATE very useful. For several participants GATE didn't identify any chemicals and for others it only picked up a few of the very common ones. Gate also often picks up chemicals wrongly as it suggests that C and H are chemicals because these can be chemical symbols, but also picks them up when they aren't being used in that context. ChemicalTagger, as evidenced by Table I.1 in Appendix I identified many more chemicals than GATE. Participant S was impressed that ChemicalTagger had picked out the IUPAC codes of the three compounds that they had made. Commenting that "it's more interesting to see what I've made" as opposed to just seeing the common structures. Some participants in each group said that they found ChemicalTagger had pulled out the best terms in relation to their work (3 Chemists, 2 Physicists & 4 Biologists). However, with similar comments to those regarding, ChemicalTagger was also described as too generic in places, in that it picked out all common chemicals.

The results of the initial user studies elicited a requirement for automatic chemical recognition, but listing them all as tags identified concerns. In addition to stating that too many generic chemicals being pulled out wasn't that useful, concerns were also raised about its accuracy as some parts of equations were picked up as chemicals, and that it doesn't always accurately differentiate between chemical elements and chemical compounds, and isn't consistent with identifying chemicals written in different ways that refer to the same structures (e.g. NH was written as HN in a chemical structure and these two were picked up individually). Ironically the participants that raised the most concerns regarding ChemicalTagger were the chemists; this could be because their work contains more chemicals, as they have made comments such as "this chemical would be used in everything I do" (for example alumina for the electro-chemists as they use that material to polish their electrodes). These comments have shown how despite wanting chemicals to be automatically recognised, it doesn't mean that they should be used as tags, and that some PhD students value picking out more individual elements of their work rather than picking out generic terms. Equally there were comments that the participants wouldn't necessarily want to get rid of those chemical associations in certain situations, Participant J commented "I wouldn't get rid of them, if you found a contamination of something in the lab and you had to look up every time you used it", suggesting that there would be uses to keeping them as metadata but without flagging them up as main descriptions of the document.

This was highlighted in the answers given when the participants were asked how they would want to tag their documents if they were doing it themselves. The chemists raised a wider spectrum of requirements for different types of tags than the physicists and biologists but some common types of tags were requested across the different disciplines. There were some chemistry specific tags that were asked for, and the rest of the tags fell into one of two categories: scientific tags such as experiment number or type, and more generic tags such as date or year. Table 6.1 illustrates the different tagging requirements that were elicited from the participants responses.

Category	Tag	Chemistry	Physics	Biology
Domain Specific	Own Compounds	✓		
	Molecules	✓		
	Sample Number		✓	
Scientific	Experiment Number	✓	✓	✓
	Experiment Type	✓	✓	✓
	Experimental Techniques/Methods		✓	✓
	Measurements/units	✓		
	Key aims/conclusions	✓		
	Key results/findings	✓	✓	
Generic	Project Name / Number	✓	✓	
	Headings	✓		
	Date/Year	✓		
	Broad Themes			✓
	Filtered Tags	✓		✓

Table 6.1: Tagging Requirements

Additionally, the participants also suggested that in addition to tagging their documents with different types of tags, they would like to know what types of documents they are, which would also enable them to search on different categories of documents. Another theme that emerged was that participants would expect a different level of tagging if they were tagging their work for themselves, or for other people, and that similarly they would expect to search through their own work differently to how they would approach searching through a colleague's. This makes an interesting contrast to the results of the first focus groups where some of the participants didn't seem to consider putting processes in place for others to access their work either after they had left or if they were indisposed, and yet here these participants were actively considering others making use of their work.

Furthermore, comments were made about how different levels of tags would be useful at different stages of one's academic career, in that a younger undergraduate might wish for more tags and be less concerned about picking out common things because they would just be starting their academic career, and additionally undergraduate work will vary more and have less of a very direct specialism like a PhD. These themes illustrate that it's not just the tagging requirements that need to be taken into consideration here, it's also who the tagging is for. It is also clear that with the varying nature of what the PhD students wished to be tagged, that there would need to be a high level of customisation of the types of tags that were assigned. This would also facilitate different levels of tagging for different groups. Hand in hand with these tagging considerations comes searching, as one of the motivations to associate these tags with documents is to create an improved semantic search that facilitates easy searching across users work.

6.3.3 Searching

Following the tagging, the participants were asked how they would assume a search feature would work and what they would expect to be prioritised by the search; they were then asked about advanced search features. The current way the search works is detailed in Chapter 5 in Section 5.4.5.4, whereby a simple term frequency weighting is used for the main search across both the title and the main body of text, or the search can be restricted to only consider the title or the text body.

There was a disparity among the participants as to how they expected a search to work. Some participants said that they would expect the search to prioritise documents that had been tagged with the search phrase as a 'highly weighted' or 'important' tag; whereas others said that they would expect how often that term appeared in the document to be the first order of priority, and yet other participants said that they would expect to see the documents where the search appeared in the title first. This simple question in itself illustrates how varying the search expectations and needs of different PhD students can be, as even within the different disciplines there were contrasting opinions on this matter. It also illustrates how variable different scientists' ways of working are, reaffirming that how scientists organise their work is a highly personal endeavour (Shankar, 2007), and therefore how they would choose to search it is also equally personal. This links to Key Finding 1 (KF1), which states that how scientists take notes and organise their work is a highly personal endeavour.

A theme that emerged from this set of questions was that most participants did wish for a more advanced search with additional options and restrictions. Several of these are ones that already appear in Google such as using Boolean operators, searching on multiple terms and using regular expressions. Additionally, groups of participants agreed that they would want to be able to search by date (which was highlighted as a desired tag in Section 6.3.2), and that they would expect options to sort the searches by date and relevance, which are again typical searching features of Google and other search engines. There were also comments that bore some similar results to the original focus groups detailed in Chapter 4, that demonstrated that the chemists typically had more complex and varied methods of organising and searching through their data; and this was reflected in the responses given in these evaluations.

The main area that focused on a semantic search was being able to search on items that had been semantically tagged in the document and being able to filter the search by different types of tags, or drill down the hierarchy of tags. For example, searching for documents that had been tagged with experiment tags, and drilling down to the different experiment numbers. There were also desires for image searching, and the ability to search by InChIKey[1] or SMILES[2] structures to ascertain which documents contain these structures, which would also require having tagged the documents with these structures in the first place.

Similarly to tagging, the participants raised the point that the searching requirements and ways of searching would vary depending if one is searching through their own work or other peoples. As Participant AK stated, “when it’s your own work you’re never searching broadly, you’re always searching specifically”, whereas the participants pointed out that if they were searching through other people’s work or if other people were searching through theirs they would expect them to use a broader search with less refined options: “I know how to look through my stuff, but if push comes to show I doubt I’d be able to find things in other people’s log books” - Participant B. There was also agreement that it would aid with knowledge transfer and encourage research groups towards better preservation of their data and work, if these enhanced services were designed well enough that they would actually be used. The participants comments showed that they would find this type of improved searching and tagging on others work more useful rather than on their own work; as this would be work that they didn’t immediately know the order or context of. This illustrates that employing these software techniques can enable scientists to get more out of other researchers work.

These results suggest that both tagging and searching are a very personal procedure for and that in order to design a search that would fit these contrasting needs, a lot more work would need to be done to ensure that both the tagging and searching was customisable. Additionally, further work would need to be done with the tagging to ensure that the right tags were captured such that the scientists would be searching on the tags they require. They also show that the PhD students view how they use their

[1] `https://iupac.org/who-we-are/divisions/division-details/inchi/`

[2] `http://www.daylight.com/dayhtml/doc/theory/theory.smiles.html`

work and how others use their work very differently, and that in some areas they see the benefits of this type of system being more for their peers or collaborators than for them directly.

6.3.4 Markup

The participants were then asked what they thought about the mark-up and tooltips, and whether they thought they were useful or not. With regards to the tooltips the similar theme of wanting to be able to customise the information re-emerged. Some of them liked the idea of having more information show up on the hovers but wanted to be able to potentially customise that information as they felt it would be more useful to see it in their own words, and that there may be instances where their work was so specialised they would need to write the descriptions themselves. Similarly, there was a general feeling that for some of the more generic descriptions, if it was their own work they would know what the terms were that they had used and would not need a description: "If I've written centrifugation I know what it is" - Participant S. However, again it was pointed out that the more generic definitions given by the tooltips would potentially be very useful for other people reading their work, or younger students who hadn't learned as much as they had. This bears similarities to the work done by Chen et al. (2006) in the hypermedia world, their studies concluded that experts and novices exhibit different behaviour in using hypermedia learning systems, and require different levels of support. There was also a resounding agreement that participants would want the options to turn the mark-up and tooltips on and off, and be able to customise information across multiple documents.

Some participants again made suggestions that it would be useful to break the tags down into types (see Table 6.1 for the different types of tags) so that they could turn different groups of tags on and off, and be able to search on different types of tags. It was also suggested that these hovers could then link to other related work, Participant S suggested that "If it picked out my compound names and then does a tooltip that linked to all of the documents that had it in that it would be amazing" (which links back to the original user requirements FRS5 and FRS6 about making links between related notebooks and projects that were not implemented as part of this first iteration).

The responses suggested that some participants would make use of the mark-up and hovers but they would need to be both customisable, and able to toggle on and off in groups, and as a whole so that they didn't have to be used or displayed at all time.

6.3.5 System Design

The participants were then asked what they thought of the system design, and how the different types of functionality (tagging/mark-up/settings) had been mocked up to show how they would work in the document itself in Google Drive.

The feedback regarding the design and overall concept of the ELN platform was positive. The participants liked the fact that it was built on top of a familiar platform, especially the biologists. Participant AQ stated "It would work for me because it's familiar, I would rather have something where I know what I'm doing and how to deal with it and the additional tools are all you have to learn"; Participant AN agreed with "I think it's nice in that it's not frightening', and Participant AQ said "I'd be more willing to learn an extension or plugin than a whole new set of software". There was also an agreement that the additional menu items sidebars were unobtrusive, and didn't take up too much space. Although more than one participant did note that they would expect some form of user manual to explain how it all worked, and that they felt like they wouldn't automatically assume that they could get to this functionality through the added menus at the top of the document. This highlights the importance of the effort expectancy factor in UTAUT as the participants made it quite clear that they would want something that they felt was easy to use, and their positivity towards building this on top of Google Drive was its familiarity and the fact that they already knew how to use it. Additionally, this could link to the facilitating infrastructure factor as using pre-existing software would mean that the users already have confidence in how well the software was supported.

There was a general agreement that the participants liked the ability to turn the functionality on and off and would want a 'turn off all' button so that everything could be switched on and off if required. Some participants also said that they would like more panes to be shown as a sidebar with the different tags split up into types as well as showing them all together, and most participants said that they would like to see more differentiation in the colours (as some of the different tag colours were deemed to be too similar) or have the option to customise them too. One issue that was raised with regards to using this was that some participants work on their images until the last minute and would ideally want a way to embed links to their images or have their images in the document in such a way that if they clicked on them it would open it up in the program it was created in (e.g ChemDraw or Origin) so that they could edit it properly. This could be considered in the domain section of the notebook to see if any similar programs could be embedded as add-ons within the document.

Overall the way things had been laid out was well received, although some participants said that they wouldn't necessarily want to have to use Google Drive, but would be more inclined to use it if it was added on to Microsoft OneDrive or other software they deemed more secure with regards to their data. This was in part due to Google Drive not being EEA (European Environmental Agency) compliant as detailed in Section

4.4.2.5, meaning that some participants would either be against using it or would not be permitted to use it due to their industry sponsors. This was one of the reasons for creating the separate backend of Semanti-Cat and de-coupling the main functionality from a specific cloud notebook platform, such that it could be adapted to work with other similar environments. This links back to Key Finding 7 (KF7) in Chapter 4 in that some PhD students seem very attached to their paper lab notebooks and do not believe that an ELN would be easy to use or integrate with their workflow in a way that would make their life easier.

6.3.6 ELN Behaviour & Thoughts

The previous user studies conducted concluded that PhD students typically had a greater tendency to use software to write up their papers/reports/thesis rather than during their lab experiments to take notes. The participants were asked their opinion on this, and then whether they would be more inclined to consider using an ELN aimed at this stage with the features included in this prototype, and whether they thought semantic tagging/improved search would have any impact on the efficiency of their work or how much they chose to digitise their work.

The participants resoundingly agreed that currently they did all use software more to write up their papers/reports etc rather than during their experiments in the lab, although Participant AK stated that “I think it’s starting to move that way though, a lot of people are having that conversation of how do we move to store our data electronically and how do we move so that we can do write-ups in the lab”. Some participants stated that they didn’t like the idea of things going between the lab and the office, and Participant AK even has special pens that they use in the lab rather than contaminate ones from the office. Participant AQ said that they wouldn’t want to take technology into the lab even though they use the same pens inside and outside the lab; and the overwhelming response was that the participants still liked using pen and paper in the laboratory, irrespective of how much of their lab work that they digitised afterwards. Participant B stated that “this wouldn’t sway me to an ELN but would add to what we already have to make it easier to search through”, and Participant A said: “There is nothing you could do to an ELN for me that would make me use it on a day to day basis”, stating that they wouldn’t want to give up their lab notebook.

These comments show how the very term ELN can come across as trying to replace the paper lab notebook rather than providing additional support to the lab process, suggesting that ELN based software might need to be marketed as a lab management tool or knowledge management tool rather than something that looks like it’s trying to be a direct electronic replacement. Participant S described the ELN that they tried as “a replacement for paper, and it’s taking something that works already and making it harder”. This also shows that there is still a resistance to ELNs even if there are software requirements that participants have mentioned that they would like. ANT denotes that if an actor is added or removed from a network it will have an effect on the entire network, and the participants have illustrated that they feel strongly about the idea of having their paper lab notebook removed, and having an electronic lab notebook added. This also shows a resistance to losing a current actor they hold a great attachment to.

When asked if they would consider using an Electronic Lab Notebook more if it was aimed at that stage of the lab process there was a positive response, as some participants agreed that it would be very useful to be able to put everything in one place where you could link it together and search it, and that having all of their document write-ups and data together in one place even though they are all in different formats would be helpful. However, the participants had different opinions about which part of the write-up process this specific prototype software would fit into. The chemists saw it as an overriding system to store everything in one place to manage and search their variety of documents. The biologists had a similar perspective, stating that they would want to continue to use their paper lab notebook inside the lab, and that this type of software would be used to manage all their electronic documents. The physicists however viewed it as more of a literature management tool that they would want to feed literature papers into to assimilate and tag. This links back to Key Finding 8 (KF8) in Chapter 4 in that the software needs identified by the study participants weren't necessarily what they would attribute to an ELN, rather than an improvement to their current software offerings. This also demonstrates a desire for a better management system for different areas rather than something to specifically create the notes themselves.

Some participants said that the semantic tagging and searching would also encourage them to use an ELN but with the caveat that the tags and search features were actually what they wanted. Some of the physics participants stated that they would be more likely to use it on others work or their own work if they were giving it to others, rather than specifically on their own work for them. This suggests that adding a new actor of an ELNs could be better received if it was done without attempting to remove or directly replace the actor of the paper lab notebook.

When asked if the participants thought that semantically tagging the documents would improve their efficiency, there was a mixed response. Some of the chemists said that they thought it would make writing up easier if everything was tagged and linked together so that all the related material for an experiment could be easily searched for. However, the strong caveat was that this would only apply if the tags were deemed useful and met with what the participants wanted: "On these tags, no! On the tags I wanted, absolutely!" - Participant S. Some of the biologists thought that it would be useful, but it would require a level of personalisation to become fully useful: Such as personalising which terms to tag and anonymising the tag descriptions. Some of the physicists stated again that they thought the semantic tags would be more useful for other people's work than their own.

The participants were then asked if they thought semantically tagging their work or improving the search capabilities around it would encourage them to further digitise their work; this was also met with mixed responses. From the chemists, there was a desire expressed for this type of software to manage their work. Participant S stated that "it would need to be that all-encompassing project management tool, everything in one place", and Participant J and AJ expressed a keenness for the tags, stating that they would find tagging and improved searching useful. Although the general feeling was that whilst the participants could get on board with tagging and adding their own custom tags, it wouldn't necessarily encourage further digitisation, even if it would encourage using this type of software on their existing digitised work. A notable exception to this was Participant AK who said that they were looking to keep more stuff digitally and less stuff in a lab book as they felt that it was easier to store multiple copies, and mentioned their constant fear that the chemistry building might burn down and that all of their paper lab books would be destroyed. This is interesting as this conflicts with Key Finding 4 (KF4), that participants are much more concerned with backing up electronic work than paper based work; although arguably despite these concerns and awareness this participant is not currently backing up their paper based work. This fits with some of the outcomes of the scenarios discussed in the initial focus groups in Section 4.4.2.8 where the participants showed that despite being aware of the consequences of losing their paper lab notebooks, which some of the participants deem very valuable, they still do not back them up. Also illustrating that thus far out of the sets of focus group participants, the social influence of supervisors enforcing paper backups still had more of an effect than the participants concerns about their paper notes.

The biologists stated that it wouldn't necessarily encourage further digitisation, some of them took the stance that their work was already as digital as they felt they could get it, and one biology participant was quite against using solely technology for their lab write-ups, and there was an overall agreement that they all felt attached to their paper lab books, and were quite happy with the amount they digitised. From the physics focus group came a debate where some participants believed that students would need to be forced to start using ELNs from an early stage to ensure that they formed the habit of digitising their work, whereas other participants thought that if they had been forced to use one they would rebel against it. Participant Q of the biologists also stated that "if my supervisor had made me use one I would have gotten on with it... but even if I had an ELN I would still scribble on paper". This shows a level of unwillingness to change formed habits and patterns (which fits with Participant AI's comments from the initial user studies that once you've started something one way you do not want to change it); and clearly the benefits of paper and the hostile environment of the still heavily influences PhD students decisions regarding how to handle their notes. This also partially fits with the social influence element of UTAUT where in some instances the participants would have been swayed to use an ELNs if their supervisor had insisted on it, illustrating that if someone more important than them pushed for this they would accept it and get on with it, but if given their own choice it's not how they would operate. Additionally, it is interesting that the participants who were most in favour of the use of an ELN and who either tried out ELN or generic note booking software were the youngest first year PhD student, who potentially had come from a slightly younger generation that was more willing to use technology at an earlier stage, and the two PhD students who had nearly finished and had secured jobs in industry and saw the benefits of using note booking software alongside their colleagues.

The final comments that were made also illustrated the barriers to ELN software. Paper is still viewed as easier to use than an electronic equivalent and some of the participants are firmly stuck in their habits of scribbling down on paper. Participant AL stated that "For me the biggest barrier would actually be digitising the work. I wouldn't be prepared to..." And Participant Q stated that "Even if I don't have paper I will get a bit of tissue and write on that and take it out with me instead of putting it in my phone or whatever because it's easier"; this fits well with the UTAUT factor of effort expectancy. Additionally, to strengthen this factor, there is still the conception that the software out there won't be easy to use, Participant B stated "Strongly the issue isn't the software, it is the hardware, writing stuff down or jotting stuff down is way easier on paper". This links highly with the effort expectancy factor of UTAUT as participants still firmly believe that using paper in the lab is easier, and earlier comments highlighted that some perceive an ELNs as a direct replacement that looks to replace paper but make things harder for them to do, which isn't an attractive option. The participants do not believe that fully replacing their paper lab notebooks with an electronic version would make their life easier, so they are naturally against it. This also shows the attachment some of the participants still feel towards their paper lab notebooks, illustrating their importance as an actor in the lab network. These issues will be discussed further in the discussion section.

6.4 Discussion

The results of these studies lend weight to the key findings made in Chapter 4 (which all still apply in this instance) and the earlier ethnographic studies discussed in Chapter 2, whilst also adding new insight to how PhD students feel regarding ELNs and what their software needs are. A discussion of the overall findings with respect to UTAUT and ANT will be detailed in this section.

The different groups of PhD students showed some similar patterns to the original focus groups in terms of discipline characteristics, reinforcing some of the different needs per discipline. The biologists were similarly uniform in their work both in terms of the documents they produced and the way they worked and even the content of their documents. They also all used a small set of software programs (Word, Excel, GraphPad and Prism) to produce their work, and were positive about the idea of building an ELNs on top of a pre-existing cloud platform that looked familiar to them, as they didn't want to have to learn an entire new piece of software. They also by and large had more basic software needs that fitted with this proof of concept software, as with respect to writing up and handing data they mainly used Word and Excel, of which Google Drive has similar versions of the software, and one of the biologists already used Google Drive to collaborate with their co-workers in the lab. Additionally, their documents were the easiest to process and tag as they were mostly words with some figures, and the figures had extensive captions, so all of the important information could be easily extracted from the documents. The physicists had some more technical needs, they primarily used LaTeX to produce their documents, meaning that they would have a high requirement for software like this to be able to handle PDFs; and had a high use of equations and figures (mostly with lesser captions than the biologists). They were more sceptical about software's ability to successfully extract and tag pieces of information from their documents, and also with regards to writing up their documents, had a requirement for additional functionality such as handing equations. The chemists showed the greatest level of difference (similar again to the original focus groups) and had contrasting opinions about what they would want tagged and produced more varying types of documents than the other two disciplines. This shows how contrasting the different science disciplines can be, and that chemistry in particular can vary greatly in terms of software needs and approach, and backs up Key Finding 1 (KF1) about how personal an endeavour note taking is for the participants.

The prototype software evaluated as part of these focus groups was created based on some of the requirements elicited from the participants in the previous studies. The software was met with mixed reactions from the participants. The simple sounding requirement of adding tags to scientific documents proved to be anything but simple. The participants made it clear that broad tags that covered the general themes of their work were not useful to them, and that they would want intelligent tags that picked out the key distinguishing elements of their work to enable them to effectively search through it. Performance expectancy and effort expectancy from UTAUT were key factors here. The participants made it clear that they would want to feel like the tags would actually add something to their work, and made suggestions of scenarios that they felt would actively improve what they were currently doing such as tags that linked to other documents containing that tag, tags that intelligently only highlighted the chemicals that had been created by the participant themselves rather than generic ones or picking out specific measurements or samples.

Automatic chemical recognition was given as a desired feature in previous studies, and yet the chemicals identified here were met with a degree of criticism. In some areas, this was a software fault as the participants stated that some of the chemicals were incorrectly identified. However, in other areas they said that they didn't see a use in having all the chemicals they used identified as tags because some chemicals would be present in almost every document, but also some stated that they wouldn't necessarily get rid of them they just wouldn't expect to see them as tags. This suggests that commonly occurring terms such as regularly used chemicals should be stored as metadata, and the unique identifying tags should be specifically stored as tags.

Furthermore, the difficult chicken and egg situation presented itself, in that the users would only want to have the tags and make use of them if they perceived them to be useful and also if the setup was easy to use (effort expectancy), but even some of the participants themselves identified that they would need to use it and personalise it in order to see its use. Participants stated that they would want to be able to add and edit the tags and descriptions, but equally would find it time consuming if they had to do all of it, and didn't see it as something they would do on a daily basis. The software therefore needs to be able to tag well enough that the users aren't initially put off by it and have enough customisation options that they can personalise it to how they want.

Searching was also elicited as a very personal endeavour for the participants, as there was a variety in their searching methods and expectations. A common theme that came out was that the participants would be looking to search specifically within their own work rather than broadly, and that they placed value on being able to either find specific pieces of work based on a chemical structure or a piece of equipment, or being able to pull together all the different documents (including different formats of documents) for the same project or experiment. A lot of advanced search features were requested, mainly around searching for the types of tags (e.g dates, experiment types, experiment numbers, units), so a tagging system where the tags themselves had types, and searching could be done on tag type would be a) useful and b) easy to implement and expand if the search was written in such a way that it searched on tag term and type. Again, the searching requirements were based on improving their performance such that they could write-up faster if they could easily pull together all the material for one experiment in a search.

The participants also made it clear that there were some software programs that they used that they heavily relied on to create their images. The chemists mentioned ChemDraw and Origin and Quartzy (ChemDraw was the highest used molecular editing software from the chemists' software usage survey in Chapter 4). The biologists mentioned PRISM and GraphPad, and the physicists mentioned PRISM and Matlab. There are clearly some key software programs that are used to handle data and create figures across the disciplines, and some of the participants noted the usefulness of LaTeX linking to images such that the original source could be edited independently, or described the virtues of being able to click on an image in a write-up and have it open in the original software program. This is useful information for the domain layer of this software, as these are the types of software that the different domains need for their work.

The overall feedback on the proof of concept software was that the participants liked the idea of an add on to pre-existing piece of software they were familiar with, and some of the participants noted that they already used note booking software such as OneNote or Google Drive. They also mostly liked the idea of tagging and searching as long as the material was accurate enough to aid their performance expectancy, and was easy to use (effort expectancy). However, whilst some of the users were in favour of Google Drive and they liked the idea of using pre-existing software, others raised concerns about Google Drive itself with respect to security and everything being stored in the cloud. Some of the participants preferred Microsoft OneDrive as it was more secure and fulfilled the European Environment Agency Privacy Policy requirements for where they are allowed to store their data. Therefore, it is important that this software is designed in such a way that it could be added on to multiple types of note booking software rather than being heavily interlinked to one. These concerns were expected as some similar comments had been made in the original focus groups and this was one of the reasons that a separate backend was built to ensure that the main functionality was decoupled from Google Drive even if the front-end mock-ups were shown with Google Drive as the platform of choice.

6.5 Key Findings

The results of these focus groups have elicited another five key findings, which build upon the first eight key findings identified from the initial user studies. These will be summarised and discussed in this section, noting which ones build on the initial key findings.

6.5.1 KF9: ELNs are still primarily perceived as a replacement for, rather than a supplement to the paper lab notebook

Several times during the user studies it had to be explained to the participants that when the term ELNs was used, (particularly in the last set of questions where whether they would be swayed to use an ELNs under certain factors), that this did not necessarily mean stop using your paper lab notebook and use an ELNs as a full replacement. Comments were made that denoted that a participant wouldn't be swayed to an ELN but would use this software to add to what they already had, which in essence was the whole point of the question, but as soon as the term ELNs was used the participants seemed under an incorrect impression about what this meant. This again highlights the potential disruption perceived by considering displacing the paper lab notebook with an ELN in the network of actors in the lab process. Additionally, when the participants were describing how they might use this type of software phrases like 'knowledge management tool', 'knowledgebase' and 'project management tool' were mentioned, highlighting that the participants didn't see this type of software as an ELNs so much as an additional organisational tool that could help them with their currently digitised work, which fits with Key Finding 8 (KF8).

How the participants view this type of software is vitally important as perhaps a simple way of encouraging more scientists to use it would be to market it under a slightly different bracket of software. Furthermore, when disassociating the notion of replacing the paper lab notebook, and looking at the software as just another tool they could use, their responses about using it became more positive, and some clear software based needs were identified. With regards to UTAUT the participants suggested that if the functionality of this software could be improved to the standard they want, that they believed that it would fulfil the performance expectancy criteria; and that the main area they believed that it would not be easy to use, therefore not fulfilling the effort expectancy criteria was if they had to start entering everything into an electronic device rather than using paper inside the lab; also highlighting that they seem more willing to accept a new actor into their network if it doesn't involve removing or replacing their paper lab notebook.

6.5.2 KF10: Tagging and searching a scientist's work is also a personal endeavour

The participants made it very clear that they all would tag and search their work in different ways, agreeing with Key Finding 1 (KF1) and the works of (Shankar, 2007; Oleksik et al., 2014). Some common themes were requested with regards to tags, but even in that instance different participants prioritised different types of tags more. Similarly, when asked which criteria the participants would expect to be prioritised with regards to searching, the participants varied in their answers due to the nature of how they organise their work (e.g. some do not use meaningful titles so wouldn't expect presence in the title to be prioritised in a search, but others would because they title their work in a way that is meaningful to them), and various different types of advanced search features were requested. This suggests that a one size fits all approach to this problem wouldn't work. Instead the tagging and searching would need to be designed in such a way that it was customisable, such that users would have a lot of control over what they opted to have tagged in the first place. The level of customisation required suggests that machine learning could be of use here, if the tags and searches were fed into a training set for the software to learn how their users wanted their work to be tagged and how they would use the search bar.

6.5.3 KF11: Some scientists attribute tagging and enhanced searching as a more useful feature when searching through other scientists work rather than their own.

During the focus groups, a commonly reoccurring point made was some of the participants felt that the tagging and searching would be of more use to them if they were searching through somebody else's work rather than their own. Previous research conducted by Chen et al. (2006) details that novice users in hypermedia learning systems use an undirected search of trial and error, whereas an expert user will perform a directed search. The comments from participants resonate with this finding, as they described that they felt more knowledgeable about their own documents and knew how to specifically search through them, but that this behaviour would not hold true for searching others work.

When the types of tags were discussed again the participants said that they would potentially add detailed tag descriptions or add extra tags if they were handing over their work to someone else or sharing it in a group rather than doing it for them. These points were made on the basis that the PhD students felt that they were specialised in their subject and well versed in what they were writing about, and in some cases already felt like they knew how to search through their own work. It also shows however that they understand the value of knowledge sharing and leaving their legacy and that they could see the use for this type of system for work that was being passed on. Additionally, the participants pointed out that some of them felt like this type of system would be very useful for undergrads who didn't have as much knowledge as they did and who also worked on a wider variety of subjects. This range of subjects would elicit different tags, rather than postgraduate work, which can be specialised to one or two main subject areas. Furthermore, the postgraduate students explained that at their level of experience they didn't necessarily require common terms to be explained, but that they would have appreciated it when they had less experience in their subjects. Both of these themes can be looked at under the guise of 'novice' and 'expert' users, in that even a postgraduate student who is an expert in their subject area, could still be a novice in another subject area, and return to having a similar level of knowledge in that field as an undergraduate would in their fields. Therefore it is important to consider an advanced directed search for 'expert' users, and an intuitive search that allows novice users to perform an undirected search through unfamiliar work.

6.5.4 KF12: ELN reluctance is strongly linked to current hardware capabilities

This study has elicited a lot of software needs from the participants, and also shown that whilst they still use their paper lab notebooks in varying degrees, there are a lot of participants who do digitise a lot of their work and are not against using software. However, they are against giving up paper and using an electronic device inside the lab. There have been on-going comments and references to the fact that they still find paper easier to use inside the lab, to jot things down and scribble notes, and that there were multiple factors that would put them off using an electronic device in the lab, including current hardware capabilities and the fact that they didn't want to damage their hardware in the hostile lab environment.

These are adoption barriers that have been consistently identified throughout both this project and previous work surrounding ELNs, and even the participant who said that they were looking into purchasing a electronic tablet they could write directly into using a stylus that would link with their OneNote identified cost as a further barrier. Tabard et al.'s 2008 study of a tablet based ELN concluded that users found the tablet time consuming to use. Technology has advanced substantially since the initial development of ELNs and yet there is still a lack of confidence that anything out there would be as easy to use as paper, and whilst devices such as the iPad Pro have made significant improvements in this technological area, however they are very expensive and some PhD students are still unwilling to take anything into the lab they feel might get damaged. This builds on Key Finding 5 (KF5) that many lab environments aren't conducive to using electronic devices. The works of Shankar concluded that scientists still use an unusually large amount of paper, and this still holds true over ten years later, illustrating that paper still has affordances over todays hardware even though it has advanced significantly over the last decade.

6.5.5 KF13: Adoption of ELNs requires more than just good technology, it requires a change of attitude and organisation

Comments that stated that participants wouldn't be willing to digitise their work further, and that nothing could be done to an ELNs that would make them willing to give up their paper lab notebook shows how strongly the participants feel about this issue, and shows how disruptive they perceive removing their paper lab notebook from their network would be. This also links to Key Finding 7 (KF7) that scientists are still attached to their paper lab notebooks, and do not believe that an ELN would make their lives easier. However other comments illustrate that some participants believe that if they had been made to do this from the beginning then they would have just got on with it, and there were suggestions that this type of software should be given to undergraduates or first year PhD students (by participants who were very close to finishing their PhDs).

Additionally, both this study and the previous user studies highlighted an unwillingness to change a method that the participant felt already worked for them, and that they would consider changing their approach to organising their work midway through to have a high overhead (thus not improving their performance expectancy). There were also patterns of behaviour elicited such as Participant Q stating that even if they did not have paper to make a note on they would use a piece of tissue rather than entering it into their phone, and when the participants in the crystallography lab observations were questioned as to why they wouldn't consider using Google Sheets instead of an Excel file that they manually updated in person alongside each other or sent through emails, they noted that they hadn't considered doing so. This shows that some participants have gotten into a habit of working in a certain way and haven't considered changing it because they feel like it works for them, and or because they feel that it's too late to change. Therefore, targeting younger students and trying to promote further digitisation of their work from an earlier point in their career, by illustrating the new hardware software offerings and making suggestions of how they could digitise their work could play a part in changing this behaviour.

Additionally, social influence clearly has an impact on some of the participants based on their comments about how they would have used an ELNs if their supervisor made them, so supervisors and lectures promoting this way of working could also have a positive effect on how students start organising their lab work. This lends weight to Key Finding 6 (KF6) which describes how some participants opinions of ELNs were influenced by their supervisors. Additionally, it's important to make clear that they are simple ways of digitising work that are at least taking a step in the right direction. For example, students could take pictures of the important entries in their lab books or of experiment results and sync them to Google Drive (if they were using that for their notes) and then when they came to writing their reports they would already have the important paper notes saved as pictures alongside their documents. Digitised doesn't always have to mean fully typed out, a picture of a lab notebook page is still better than it only existing in paper form, and still provides more of a backup and improves access to material outside the lab. These simple changes need to be both considered and promoted to make slow progress to further digitising the scientific record. Some of the participants stated that they felt that ELNs and a digital lab environment would be a change that would happen, but just not yet, and making some of these slow changes could spur that change forward.

6.6 Conclusions

The initial conclusions of the previous user studies were that an ELN would need to be created that didn't try to replace paper and that aimed to improve their current software needs regarding their already digitised work, and that scientists weren't against the use of technology they were against the idea of removing their paper lab notebook. The findings of this study agrees with those conclusions, the discussions around the software needs of the participants highlighted that there is new functionality that is desired by scientists and there is a wish for a better way to organise and collate their data. This suggests that despite resistance to ELNs scientists aren't all against the idea of managing their work digitally. The scientists seem willing to do more with their already digitised data, and are looking for better ways to manage it, and indeed participants made it clear that they viewed this piece of software as more of a knowledge management tool; showing that this is where they see a potential performance improvement. However, when the notion of removing their paper lab notebook was mentioned, and when they considered using ELN based software as a replacement for their paper lab notebook rather than an additional helpful tool, that was where a level of resistance was identified.

This shows that depending on how one views an ELNs, some scientists aren't actually against ELNs they are against the idea of having to give up their paper lab notebook. This shows the very strong attachment scientists have to their paper lab notebook and to the affordances of paper, and that they feel very strongly about the removal of their paper lab notebook. The paper lab notebook is therefore clearly a very important actor that scientists perceive as causing a great disruption if it was removed from their network, and that removing it would not improve their performance expectancy. These findings fit with the initial conclusions derived from the initial user studies, both with respect to how the paper lab notebook is viewed and how important the UTAUT factors of performance expectancy and effort expectancy are. An ELN marketed under the bracket of a lab management tool, that targets improving their performance with regards to organising and managing their work and being able to search through it more accurately could have more success.

Chapter 7

Conclusions

This section describes the conclusions that have been made with respect to each research question, detailing which research objectives were completed for each question. It then proposes four different potential avenues of future work that could be undertaken to continue or complement the work completed in this thesis.

7.1 Conclusions

This section begins with the overarching conclusions that have been drawn from the work conducted in this thesis, including the socio-technical conclusions that have been formed that lie outside the remits of the research questions. Following this, the conclusions derived for each research question will be answered, stating which research objectives were completed to answer each question. Research Question RQ is the Primary Research Question, and the other research questions are subsidiary questions designed to elaborate on and inform the answers to these questions.

7.1.1 Overarching Conclusions

We are still not in a place to fully replace paper; scientists still regard its affordances very highly and there are still many barriers in place such as the hostile lab environment, that mean that some scientists are unwilling to use an electronic device in the lab. However, despite this there is still clearly a need for additional software support. None of the scientists involved in the studies said that they were perfectly happy with their current software offerings, and there were many comments made about how things could be improved. There is a gap for a piece of software that functions as an overall management tool to manage all of the scientist's digital work, and perhaps if scientists can be shown the benefits of having all of their digital work in one place it might encourage them to digitise further.

There is also a stigma attached to the idea of ELNs, which was illustrated in the software evaluation focus groups when participants reacted strongly to the idea of using an ELN, but when explained that in this context it was aimed more towards the write up stage of the lab process and did not aim to replace their paper lab notebook, they were more in favour of the idea. A lot of the functionality that could be provided by ELN software is recognised as valuable by the scientists, and yet it receives more positive feedback when marketed as a organisational tool rather than a direct paper replacement; as evidenced by the participants describing the prototype Semanti-Cat as a 'knowledge management' tool. This demonstrates the extent to which replacing the paper lab notebook would disrupt the current working practices and how this isn't an issue that can be purely solved with software, and that trying to do so won't be successful.

Additionally, there are more barriers to the adoption of ELNs than merely hardware and software. Rather than focusing on creating new pieces of software to replace the paper lab notebook, we should be focusing not only on improving the current software needs where we can, but also in considering how these adoption barriers can actually be mitigated. In order to increase the digitisation of the scientific record and improve the adoption of ELNs scientists need to be caught early so that their patterns of working can be started off using more technology, as a majority of the participants showed an unwillingness to change how they currently worked. Supervisors and others who are in more important roles could also play a part in helping shape how scientists work, to ensure that they start off with good habits such as regular backup procedures, and ensuring access to their students work. Finally, if scientists still consider that using an electronic device in the lab would be an imposition and a high potential for device breakage or contamination, perhaps other solutions need to be considered when approaching data entry in the lab.

7.1.2 Primary Research Question (RQ): What influence would a cloud based semantic laboratory notebook have on the digitisation and management of scientific research?

The conclusions to this research questions were derived from working on Research Objectives RO7 and RO8. Based on the software evaluation focus groups for the proof of concept software, which included both an evaluation of the software and a discussion around the concept of the cloud based semantic laboratory notebook, the main conclusion was that this would have a much greater influence on improving the management of the scientific record rather than vastly furthering the digitisation of it. The study participants said that they saw this type of software as more of a knowledge management or organisational tool, and expressed a desire for a tool such as this as long as it actually provided useful functionality. However, the participants mostly agreed that this tool wouldn't necessarily entice them to further digitise their work, as some of the participants were against the idea of digitising their work more than they already do, and others said that they thought their work was already digitised to an appropriate state, or as much as it could be.

Furthermore, in line with Key Finding 7 (KF7) scientists are still attached to their paper lab notebook. The extensive user studies illustrated that there were many adoption barriers to adopting an ELN that still hold true, including the disruption it would have to current working practices, and the hostile lab environment towards technology. Therefore, this clearly isn't just a software issue (as demonstrated by Key Finding 12 (KF12), and software in itself could not mitigate these barriers. However, a significant need that was elicited from multiple user studies was that a lot of the scientists' work (even all their digital work) isn't cohesively organised, and is stored in multiple different formats and locations using different software and as of yet there isn't an overarching platform that allows them to collate all of these records together in a useful manner. Adding semantic tags of different types to tag and categorise different documents would enable scientists to link their documents to their data and to easily find material related to one experiment to facilitate an easier write-up.

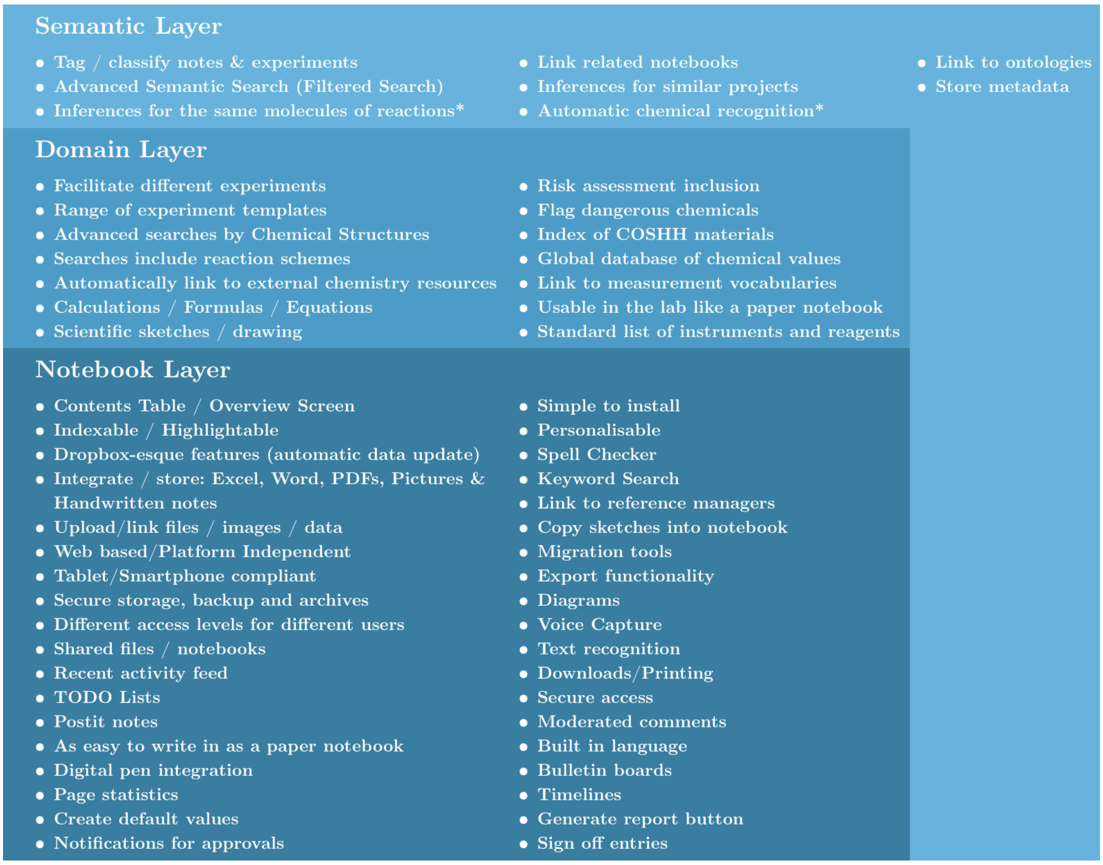
For a majority of the participants in the software evaluation focus groups it seemed like the management side of things was where they saw a benefit to this type of software. Many participants intimated or directly said that this wouldn't encourage them to further digitise their work, even though they would consider using it on their already digitised work. These participants were all but one of the biologists, one of the physicists and half of the chemists. However, others were more in favour of digitising their work and had actively been taking steps to do so. Interestingly the participants who showed a higher proclivity towards digitising their work were two physicists who were nearing the end of their PhDs and had started working in industrial labs, and a chemist and biologist both of whom had supervisors who were in favour of electronic lab notebooks (although the chemist was nearing the end of their PhD and the biologist was a first year PhD student). The rest of the chemists sat in the middle of these opinions, and weren't against the idea of further digitising their work but were more skeptical that any system could provide them with the requirements they would actually want. The discussions in the focus groups suggested that something more than just a new piece of software would be required to either consider removing paper from the lab or to drastically increase the amount that scientists digitised their work, such as starting these types of practices earlier and for supervisors and employers to set them on this road as early as possible to ensure that they get into good habits from the start.

This links back to Key Finding 13 (KF13) that adoption of ELNs requires more than good technology, it requires a change of attitude and organisation. The social influence factor of UTAUT illustrated the importance of the social hierarchy in the lab, demonstrating how a PhD students supervisor had the capacity to influence not only their opinions and attitudes towards ELNs, but also their working practices. The PhD students who had supervisors that favoured the use of ELNs were more positive about their usage as opposed to the PhD students who had supervisors who were against them, and the PhD students who had the best backup practices for their paper notes had been heavily influenced to do so by their supervisor. Furthermore, considering the principles of ANT, the PhD students were concerned at having their paper lab notebook replaced by an electronic lab notebook because of the disruption it would cause to their current working practices, however as evidenced by the social influence, some of these practices had been put into place by the students supervisors or superiors. Therefore some of the responsibility falls on the supervisors of PhD students and research groups to instantiate good backup practices, exercises where students make sure they not only digitise their work but make it available to colleagues and supervisors where appropriate, and also to try and create an environment that works better with the use of technology.

7.1.3 RQ.1: What are the approaches and features that should be taken into account when creating an ELN?

The conclusions to this research questions were derived from working on Research Objectives RO1, RO2, RO3 and RO8. With regards to the approaches that should be taken into account, this project concluded from previous literature and published ethnographic studies that user centred design, ethnography and collaborating with domain experts are all useful approaches to take. These all hold similar themes, regarding involving the users of the ELN in the design process and the studies that implemented them such as Oleksik et al. (2014); Hughes et al. (2004a) noted that these approaches did produce more useable systems that helped better understand the context in which the lab book is viewed. With respect to the features that should be included in an ELN, the literature research and user studies the features desired by users have been collated into Figure 7.1 which shows the three layered approach to a new ELN platform.

This model was originally conceptualised in Chapter 2 (see Figure 2.2 in Section 2.6.5) where the features that should be incorporated into an ELN suggested by literature were organised into these three layers. The notebook layer denotes the generic notebooking features that could be found in any Electronic Notebook software, such as spell checking, linking reference managers and creating contents pages. The Domain based layer contains functionality specific to the discipline of the ELN, in this instance chemistry such as linking to COSHH materials and drawing scientific structures. The semantic layer contains the features that use semantic web technologies such as marking up the documents, using an advanced semantic search and linking to ontologies. The user desired features that were elicited from the Dial-a-Moelcule studies and the initial user studies of this project fitted into this model and were split up accordingly.



Semantic Layer

- Tag / classify notes & experiments
- Advanced Semantic Search (Filtered Search)
- Inferences for the same molecules of reactions*
- Link related notebooks
- Inferences for similar projects
- Automatic chemical recognition*
- Link to ontologies
- Store metadata

Domain Layer

- Facilitate different experiments
- Range of experiment templates
- Advanced searches by Chemical Structures
- Searches include reaction schemes
- Automatically link to external chemistry resources
- Calculations / Formulas / Equations
- Scientific sketches / drawing
- Risk assessment inclusion
- Flag dangerous chemicals
- Index of COSHH materials
- Global database of chemical values
- Link to measurement vocabularies
- Usable in the lab like a paper notebook
- Standard list of instruments and reagents

Notebook Layer

- Contents Table / Overview Screen
- Indexable / Highlightable
- Dropbox-esque features (automatic data update)
- Integrate / store: Excel, Word, PDFs, Pictures & Handwritten notes
- Upload/link files / images / data
- Web based/Platform Independent
- Tablet/Smartphone compliant
- Secure storage, backup and archives
- Different access levels for different users
- Shared files / notebooks
- Recent activity feed
- TODO Lists
- Postit notes
- As easy to write in as a paper notebook
- Digital pen integration
- Page statistics
- Create default values
- Notifications for approvals
- Simple to install
- Personalisable
- Spell Checker
- Keyword Search
- Link to reference managers
- Copy sketches into notebook
- Migration tools
- Export functionality
- Diagrams
- Voice Capture
- Text recognition
- Downloads/Printing
- Secure access
- Moderated comments
- Built in language
- Bulletin boards
- Timelines
- Generate report button
- Sign off entries

FIGURE 7.1: Revised Feature Model, adapted from (Kanza et al., 2017)

The semantic requirements that were implemented have been given further descriptions based on the feedback of the final focus groups, and the rest of the requirements have also been expanded on based on the feedback. Below are the expanded descriptions for each of the requirements in the semantic layer, from right to left in Figure 7.1.

7.1.3.1 Tag / classify notes & experiments

This requirement was implemented in Semanti-Cat and the feedback made it clear that the tagging needs to be much more specific and refined. Rather than tagging everything associated with the document, some additional natural language processing will need to be performed over the document to not only pick out key terms, but to then ascertain which are unique enough to be used as tags. Additionally, new tagging services need to be explored to pick out less domain specific terms such as dates. Furthermore, the documents themselves should be classified as different types of documents (e.g experiment write-up, literature review) so that different types of documents can be searched on (which is similar to how many scientists already organise their paper based work).

7.1.3.2 Link Related Notebooks

This requirement wasn't implemented in Semanti-Cat due to it's dependancies. The markup and tagging would need to be successfully implemented on a per notebook basis before linking related notebooks could be considered. However, the feedback given in the software evaluation focus groups provided enhanced details on what should be considered when implementing this requirement. Linking to related work was mentioned in the final focus groups, and many of the participants were pro collaborative activities, and indeed looked at the tagging as something that might be more useful for other people looking at their work. However they made it clear that they would want to personalise their tags, so there would need to be ways of linking types of tags together across different notebooks whilst still allowing different users to edit their own tag descriptions. Another main criterion for linking that was given was linking together experiment material to make write-ups easier, so this should also be considered in this feature.

7.1.3.3 Link to Ontologies

In Semanti-Cat the available ontologies that were small enough to be loaded into Jena were used. The feedback from the focus groups was that more ontologies need to be linked, and only the discipline specific ones should be used rather than applying all of them. Some initial tagging of broader themes needs to be performed (by OpenCalais or something similar) to identify the main discipline and then only pull out terms from those related ontologies. Additionally, only terms a certain level down a hierarchy should be used as tags, as the users said that a lot of what was pulled out of the ontologies were too generic as the parent's terms were often also pulled out, and that they wanted more discipline specific terms.

7.1.3.4 Advanced Semantic Search (Filtered Search)

In Semanti-Cat a very basic search was implemented with the logic that the initially implemented features such as tagging and linking to ontologies would need to be successfully implemented first to ascertain how to design the search. The feedback about searching was that a semantic search would need to be customisable so that the different types of tags could be searched on. Additionally, the 'Advanced Search by Chemical Structure' domain based feature was highlighted as something that chemists would want to search on, therefore this could be included as part of the overall search rather than being its own feature. Furthermore, given the different ways that the participants search, a search that uses machine learning to learn how the users search, by training over their initial searches and tags could help to make a user's individual search more aligned with their own searching methods.

7.1.3.5 Inferences for similar projects

Similarly to implementing linking related notebooks, this requirement wasn't implemented in the first version of Semanti-Cat as the tagging would need to be successfully implemented to inference similar projects. In the feedback from the focus groups, a key factor identified was identifying projects or experiments where the same materials or equipment had been used, so this should be taken into consideration for making these inferences. Some participants in the software evaluation focus groups stated that they would like to be able to hover over chemicals or specific tags and see other projects that include them; similarly a desire to be able to search for all documents and data files related to the same project was also mentioned as part of these discussions. These pieces of functionality would require the ability to recognise and infer if documents belonged to the same project.

7.1.3.6 Store Metadata

In the initial implementation of Semanti-Cat stored all the tags and chemicals were stored as metadata. However, based on the feedback from the software evaluation focus groups, some of the chemicals could just be stored as metadata with the tags and other chemicals stored as metadata but also publicly exposed as tags. The participants stated that that they wouldn't get rid of these terms but they wouldn't expect them to be stored as tags against their own work; although seemed more in favour of adding more tags and descriptions if they were giving their work to others, suggesting that there should be different levels of tagging and descriptions depending on the user's familiarity with the document content.

7.1.3.7 Inferences for the same Molecules of Reactions

In Semanti-Cat the initial requirement of identifying chemicals was implemented, as in order to inference between molecules of the same reaction, first the identification of molecules needed to be successfully implemented. During the software evaluation focus groups, a participant raised the issue of understanding molecules, whereby the tags had failed to identify the family of molecules that they used. However, no further information about this specific requirement was elicited.

7.1.3.8 Automatic Chemical Recognition

The chemical recognition that was implemented in Semanti-Cat needs some work, GATE wasn't well received with regards to its ability to recognise chemicals. GATE only *tags compound formulas (e.g. SO2, H2O, H2SO4 ...) ions (e.g. Fe3+, Cl-) and element names* *and symbols (e.g. Sodium and Na). Limited support for compound names is also provided (e.g. sulphur dioxide) but only when followed by a compound formula (in parenthesis or commas)*[1]. ChemTagger tags both chemical elements and compounds, and recognises many more chemicals than GATE, thus it should be focused on with additional natural language processing to refine how chemicals are recognised, and extra functionality should be added so that widely used chemicals are stored as metadata not tags.

7.1.3.9 Overall Conclusions

Overall quite a lot of the features in Figure 7.1 were reinforced by the last set of focus groups, illustrating that these are the types of features that users want. However, it's also made it clear that for some features such as tagging a lot more work needs to be done to improve how these features are implemented, and that a lot more can be elicited from users if they are given a physical piece of software to evaluate rather than being asked what they want as an abstract notion, thus reinforcing using ethnography and user centred design as part of the approaches to creating ELNs in the first place.

7.1.4 RQ.2: What are the key processes of digitising scientific research?

The conclusions to this research questions were derived from working on Research Objectives RO3 and RO8. The main findings for this question came out of the focus groups and lab observations. There were some commonalities across disciplines despite the participants storing different pieces of information in their lab book. Typically participants from all three disciplines would plan their experiments in their paper lab book, record their observations in their paper lab book during the experiment, and then write up the necessary experiments as part of their papers / thesis / reports. In the case of the chemists and physicists, typically the data from their experiments was already digital as they generally used machines linked to pieces of scientific equipment that would produce digital data, although the biologists didn't follow this pattern.

However, an interesting observation that came out of considering this research question was noting what elements of the lab didn't get digitised, and which ones did. Some of the participants stated that the pieces of work that they didn't digitise were things that were only useful in the moment or a list of things that didn't work; which could end up being work that doesn't quite work or look interesting at the time, but may still be useful at a later date in other experiments. Therefore the PhD students would have a large set of notes to go with each experiment or study, including previous attempts at experiments or earlier versions of protocols. However, these pieces of information didn't make it to the final reports, or research papers, as typically they would be looking to

[1] `https://gate.ac.uk/sale/tao/splitch23.html#sec:parsers:chemistrytagger`

write up and publish their perfected protocols or the versions of the experiments that worked.

All of the participants created formal documents where their data was displayed in its final analysed form, and their experiments and results were written up, and a majority of the data collected was either already digitised in its raw format because it had been created by a machine and written to a computer, or it was input into a computer for processing/creating figures. However, the scribbles and jotted down notes of the lab book, such as this protocol didn't work, or this method works slightly better here, comments that have been added to printed out literature articles, and raw workings out of equations, and indeed any rough working that lent itself significantly more to paper than a computer were the pieces of information that risked remaining only in paper form.

There was a mixed reaction between the scientists about how important this un-digitised information was; some participants made it clear that they would be horrified if they lost their lab book, whereas others said that they didn't believe that it would take them long to reproduce the material in there and that it would be the loss of their samples and equipment that would really impact their work. However, these informal notes could still be important, both for the purposes of patents to show early working, and also if any mistakes are discovered in the later digitised results it would be useful to go back to the initial workings. Furthermore, noting down what doesn't work can be equally important as noting down what does work, not just for the scientist in question but also for anyone they are collaborating with or who may take over their work at a later date, so that the same mistakes aren't repeated.

Additionally, quite a bit of the work written down on paper were links to the digital data, which didn't get digitised because they were in the lab book to show the participants where to find the related data files. However, given that one of the uses suggested for the proof of concept software was a place to store all the scientists' data in once place and be able to easily search for all of the related material for one experiment, this suggests that some of these links do need to be digitised, which is where the proof of concept system and the semantic tags that enable these documents to be pulled together in a search would be really useful.

7.1.5 RQ.3: What is an appropriate cloud environment to create an ELN with?

The conclusions to this research questions were derived from working on Research Objectives RO4, RO5 and RO6. Google Drive/Google Docs was chosen as the platform to create this ELN. Initially five different cloud notebooks were trialled against the non-functional requirements (from Section 5.1.1 that were taken from the user desired features that specified operational criteria rather than specific functionality, that were elicited from the initial user studies in Chapter 4. The prospective cloud notebooks were also examined against the the adoption barriers (from Table 2.1 in Chapter 2) and whether it was free to develop add-ons to their environment or whether software/development licenses were required. OneNote, Google Drive and Evernote were identified as the three that fulfilled all of these criteria. These three notebooks were then mapped against the notebook functional requirements identified in the bottom layer of Figure 7.1, and Google Drive fulfilled the most requirements so this was chosen.

The experimental coding that was conducted as part of RO5 demonstrated that it is possible to create add-ons through various means to expand and enhance the Google Docs environment using Google App Scripts, Preliminary investigations successfully proved that it is possible to add in third party services such as ChemSpider (RO6) to a Google Document's web interface, using Google App Scripts ability to serve external HTML services, and it's also possible to manipulate the documents content, add structure, and unobtrusively link to add-ons with custom menus and sidebars. Some disadvantages were identified to the programming side of this environment, such as security and cross-site scripting errors. Some participants noted that they were not allowed to use Google Drive or Dropbox by their industry sponsors, and that Google Drive was also not permitted on some projects due to it not conforming to the European Environmental Agency Privacy Policy. These disadvantages combined with the comments made in the initial focus groups about security concerns regarding services like Google Drive and Dropbox, led to the decision of creating the backend as a separate web application that used the Google Drive API to decouple the semantic functionality from the platform. Google Drive however was used for the mock-ups to illustrate how this functionality could be put into practice in electronic notebook software.

7.1.6 RQ.4: Where could an ELN fit into the current lab practice where it would actually be used?

The conclusions to this research questions were derived from working on Research Objectives RO3 and RO8. The results of the initial user studies suggested that there was a place for an ELN or for additional software support in the write-up stage of the lab process. Scientists do use technology and do digitise varying degrees of their work, but typically a lot of what gets digitised are the more formalised documents such as reports, presentations, posters, papers and in the case of the participants of these studies PhD documents. These findings were backed up by the software evaluation focus groups where all the participants strongly agreed that the write up stage of the lab process was where they most used software, and many of the participants viewed this type of software as knowledge management or organisational software rather than a direct replacement to their paper lab notebooks. Additionally, a lot of the participants across both sets of focus groups and lab studies made it clear that they had no intention of giving up their paper lab notebooks anytime soon, and that irrespective of what work they digitised, they still felt that for the initial jottings down and scribbles of the lab, paper could not be replaced.

7.1.7 RQ.5: How can Semantic Web technologies be utilised in this ELN environment to improve the human laboratory interaction?

The conclusions to this research questions were derived from working on Research Objectives RO6, RO7 and RO8. Findings from multiple studies both conducted and analysed as part of this project has highlighted that the ability to improve searching through a scientist's vast array of documents in many formats is very useful and desired. Improved ability to search was totalled as the top priority for the users in the Dial-a-Molecule survey regarding the features users wanted from ELNs. Additionally, in the initial focus groups the participants demonstrated how they already categorised their work and made their own organisational structures for how they wanted to be able to link their work together, any many of the participants described how they used codes or other pieces of information to link between their paper and digitised work. Semantic web technologies can provide these links and facilitate a more advanced search. Whilst the tagging that was trialled in the proof of concept system was met with mixed reactions, the participants expressed interest in the idea of tagging their work and making it more searchable, they just wanted more refined tags. They made it clear that they saw a use in being able to organise and manage all of their work in once place and easily search for all of the material on one experiment or project, or being able to search very specifically throughout their work would be very useful. With further work on the tagging and searching, potentially incorporating some machine learning techniques to train a system to learn how the different users tag and search semantic web technologies could be used effectively to aid with the organisation and management of scientists' records.

Additionally, a lot of the scientists said that they felt that semantically tagging work and adding descriptions and mark-up would be very useful for collaborative work or for processing other scientists work to immediately gain more information about it, therefore semantic web technologies could also be used to improve how scientists work with each other and transfer knowledge.

7.2 Future Work

There are several avenues for future work that have been identified for this project. These are detailed below:

1. User studies with Hybrid Notebooks

2. Industry studies
3. Paperless Labs studies
4. Semanti-Cat Development

7.2.1 User studies with Hybrid Notebooks

The outcomes of this project have illustrated that there are many affordances of paper that still entice scientists to use it for several note-taking endeavours as opposed to using an electronic device. Since this project has concluded that currently we are not in a state to replace paper, the first avenue of future work could be to conduct user studies to understand how a hybrid notebook could work, where paper and technology can be used together.

It would be interesting to explore different methods of increasing digitisation through unconventional methods such as taking photographs of lab pages and automatically saving them to note booking software such as Google Drive or OneNote. Simple software could be written to automatically organise lab notebook pages into dated folders, which would only require the users to take a photograph of each lab page using a phone app. This could then link to the users note booking software of choice such that whilst writing up their work they could easily access photographs of their lab books.

Alternatively, Smart Paper and Pens Systems such as Bamboo [2] could be trialled. This system provides special surface for users to place their paper notebooks on, and a special pen to write on the paper and save their notes to a computer via a mobile app. This would allow users to use the affordances of paper notes whilst still allowing them to be digitised.

Trialling these different methods that do not aim to replace paper, but aim to work with it to improve the digitisation of the scientific record could lead to new strategies towards ensuring that the scientific record is digitised and maintained for prosperity.

7.2.2 Industry studies

Another avenue of future work would be to conduct some of the same studies conducted in this project in industry. The initial focus groups and lab observations could be conducted in industry to gain further understanding of how industry and academic lab environments differ (and which elements are the same), and what their attitudes to ELNs are.

It would be interesting to see whether the note taking practices in industry differ from academia, and if they differ across different industry environments. Furthermore, for

[2] `http://bamboo.wacom.com/introducing-bamboo-spark-write-on-paper-save-digitally/`

industry environments where the lab practice is mandated by the company practices, how their employees feel about these practices and adhere to them when they potentially lack the freedom of the academic environment could help shed light on how putting stricter practices in place would play out.

Another area of investigation that could be done as part of the lab observations is looking at the hardware and software that the industrial labs use. The lab observations at the University of Southampton demonstrated the use of software packages that do not integrate well together, and computers that used old operating systems and didn't produce data in useful formats. Comparing this to industry could produce insights into how different hardware and software offerings impact on the efficiency of lab work, and how scientists perceive the use of technology in the lab. These insights could also help identify where practices are more efficient in industry and academia and make recommendations about how one area could learn from the other.

7.2.3 Paperless Labs studies

There are some industrial establishments that have converted to paperless labs such as Lonza, a Pharmabiotechnical company [3]. There are many studies of paperless labs that could be done as a future avenue of work for this project.

A review of the companies that have paperless labs, which tools they use, and statistics about their companies such as size, country, tools used and domain could provide some valuable insight about which companies have been able to make this lab process work.

Another interesting avenue would be to interview members of these companies to gain further insight into how they transitioned into these paperless labs, and to understand the motivations behind this decision. Having elicited from the user studies of this project that some pieces of the scientific record that do not get digitised are things that do not work, it would be interesting to question users and managers of paperless labs to see how much of the previously digitised work has proved useful later in other circumstances.

Furthermore, it would be useful to observe these labs and see how they differ to the paper based labs that have been observed as part of this project. Discussions could be initiated with participants to understand how efficient they find the paperless lab process, and observations could be made to understand how the paperless lab process works during experiments.

[3] `https://www.lonza.com/about-lonza/knowledge-center/events/pharmabiotech/webinar-business-case-for-a-paperless-solution-2017-04-25.aspx`

7.2.4 Semanti-Cat Development

The final avenue of future work is to continue the development of the prototype Semanti-Cat. This development can be broken down into the following four stages.

1. Further iterative software development of the semantic layer
2. Iterative development of the domain layer
3. Packaging the semantic software as one web service
4. User evaluation studies of the whole software

7.2.4.1 Further iterative software development of the semantic layer

The feedback from the user studies made it clear that there was potential for this semantic layer but that it would need a lot more work, and the next iteration of this software should incorporate this feedback. For the ontologies, further ontologies should be searched for and added, and OpenCalais with some additional natural language processing should be used to identify the main discipline to then only use certain ontologies with. Similarly, further natural language processing should be done to narrow down the OpenCalais tags pulled out, and other tagging services that pull out more general content such as dates should also be evaluated to pick the best one to use in this system. With regards to the chemical recognition, some further processing should be done to store the regularly used chemicals as metadata and keep any rarely used chemicals across the document corpus or new chemicals as tags. The tags should also be broken into different types such as date, experiment, project etc, and based on the tags and the natural language processing the documents should also be categorised. Whilst GATE didn't produce successful results with respect to tagging chemicals, there are other elements of natural language processing available in its system, so these could be worth looking into with respect to some of the other types of tags. Additionally, the co-occurrence of terms that was computed as part of the initial development of Semanti-Cat can be studied further to understand what terms commonly occur with other terms across a corpus of documents to add further context to categorising the documents.

The search should be built in such a way that the different types of tags and documents can be searched for. Further research should be done into potential systems to train this type of data so that the users tags and searches could be fed into a machine learning system to learn how the users tag and search their data. The mark-up of the tags should be made significantly more customisable such that different types of tags can be turned on and off, and the hovers/descriptions should also be made customisable.

7.2.4.2 Iterative development of the domain layer

The features of the domain layer should also be implemented in add-ons. The feedback from the software survey and both focus groups should be used to determine the popular types of software and chemistry resources that should be considered for these add-ons. Molecular editing tools were listed as a very popular type of software in the software survey, and ChemDraw was both identified as the most popular software in this area in the survey, and mentioned consistently throughout the focus groups. ChemDraw however is a proprietary piece of software, but Avogadro, the third most popular piece of software in this category is open source, so an add-on for this could be experimented with, or a further comparison study to ChemDraw could be performed to see if scientists would use Avogadro if it was implemented within a document environment or if ChemDraw was significantly more popular. If that was the case then ways to embed links to images in documents could be investigated to ensure that they could easily be edited in their main software environment. Another of the domain requirements was to link to external chemistry resources, a number of chemistry databases proved to be quite popular from the software survey so linking to them should be investigated. This domain layer should be created with standalone add-ons, which can then be ported into Google App Scripts to make Google add-ons that can be linked to from custom menus and installed in a user's documents.

7.2.4.3 Packaging the semantic software as one web service

The semantic layer should be refined and packaged as one service that can be sent a document and will send back the different types of tags, and descriptions. This will not only maintain the stance of decoupling the functionality from the notebook platform, but also means that this service could be used by any notebook platform where the documents could be extracted and sent over. This software could also be broken down into specific domain versions that purely focused on one discipline, as well as the main one which catered for more than one discipline, depending on a user's preference.

7.2.4.4 User evaluation studies of the whole software

Once the domain layer has been developed and the semantic layer has been improved, this software can be trialled and evaluated again. This time a more extensive user study should be conducted, ideally several studies would be conducted for participants to extensively test out the different features of the software, both domain and semantic, to see how the domain based add-ons fared with regard to actually writing up their work, and how well the semantic layer tagged them. Following that some focus group discussions should also be conducted to evaluate, in groups, how well the implemented functionality meets their requirements.

Appendix A

ELN Vendors

The full spreadsheet of ELN Vendors can be found in the pure dataset https://doi.org/10.5258/SOTON/D0384 under the file name ELNMarketStudy.xlsx.

A.1 Inactive ELN Vendors

Vendor	ELN	Notes
Accelrys	Accelrys ELN	Merged with Symyx in 2010. http://accelrys.com/about/news-pr/merger-0610.html Acquired Contur in 2011 (which merged with Elevate). http://accelrys.com/products/eln/contur/announcement.html Acquired and merged with VelQuest in January 2012. http://www.cambridgenetwork.co.uk/news/accelrys-acquires-velquest-corporation-for-35-million/ Acquired by Dassault Systems in April 2014. http://www.3ds.com/press-releases/single/dassault-systemes-successfully-completes-acquisition-of-accelrys/
Amphora Research Systems	Open ELN	ELN was discontinued in 2015 http://www.limswiki.org/index.php/Amphora_Research_Systems,_Inc.#_note-OELNProdArch-3
Array Genetics	NucIt	Doesnt meet CENSA Standard (Rubacha et al., 2011)
ArtusLabs	Ensemble ELN	Acquired by PerkinElmer March 2011. http://www.perkinelmer.com/aboutus/pressroom/pressreleasedetails/articleid/69844
Cambridge Soft	E-Notebook	Acquired by PerkinElmer in March 2011. http://www.perkinelmer.com/aboutus/pressroom/pressreleasedetails/articleid/69844

Cognium Systems	iPad	Dissolved in Summer 2014. http://www.limswiki.org/index.php?title=Cognium_Systems_SA
Contur	CounturELN	Changed ELN from ConturELN to iLabber in July 2011. http://www.limswiki.org/index.php?title=Contur_Software_AB1 Merged with Elevate - LifeDoc. Acquired by Accelrys in 2013. http://accelrys.com/products/eln/contur/announcement.html
EKM	LABTrack	Used to distributed LABTrack then had legal disputes with LABTrack. http://www.limswiki.org/index.php?title=LIMS_vendor
Elevate	Life Doc	Merged with Contur. (Rubacha et al., 2011)
Elsevier MDL	MDL Notebook	Merged with Symyx. http://newsbreaks.infotoday.com/NewsBreaks/Elsevier-to-Sell-MDL-to-Symyx-Technologies-37260.asp
E-nnovate	e-notebook	Discontinued (Rubacha et al., 2011).
Evernote	Evernote	Doesnt meet CENSA Standard (Rubacha et al., 2011).
EZQuant	EZQuant-ELN	Website is non existant.
Infotrieve	Infotrieve ELN	Discontinued (Rubacha et al., 2011).
Identic Software	Invent	The website referenced in (Rubacha et al., 2011)'s study no longer exists (domain is up for sale).
Irisnote	Irisnote	Originally Recentris developed the CERF ELN, and the company was then absorbed into Irisnote Inc https://www.limswiki.org/index.php/Rescentris,_Inc.. However, in 2013 irisnote shut down and Lab-Ally LLC took over CERF Notebook.
Knowligent	Research Notebook	Website is non existant.
Laboratory Data Solutions	Labnotes	Dissolved in September 2013. http://wck2.companieshouse.gov.uk//wcframe?name=accessCompanyInfo
Labtronics	Nexxis ELN	Acquired by PerkinElmer in May 2011. http://www.perkinelmer.co.uk/CMSResources/Images/44-127929PR_2011_05_16.pdf
Macs in Chemistry	Lab Notebook	Reference website given in (Rubacha et al., 2011) doesnt exist anymore.
Mettler Toledo	Virtual Lab	Doesnt meet CENSA Standard (Rubacha et al., 2011).
Neudesic	Neudesic ELN	Company doesnt seem to market this product anymore. https://www.neudesic.com/

Open Source Project	Electronic Laboratory Notebook	Funding ended in Winder 2007, development was discontinued. `http://collaboratory.emsl.pnl.gov/software/software-notices.html`
Open Source Project	LabJ-ng	No releases since 2013. `http://sourceforge.net/projects/labj/`
Open Source Project	PNN ELN	Uses Semantic Web. Development ceased in 2007. Latest OS release on SourceForce in 2008. (Myers et al., 2001)
Open Source Project	tags4lab	No releases since 2013. `http://sourceforge.net/projects/tags4lab/`
Open Source Project	The Monster Journal	No releases since 2009. `http://sourceforge.net/projects/monsterjournal/`
Recentrics Inc	CERF Notebook	Initially merged into Irisnote, Inc, and then into Lab-Ally LLC which now develops CERF. `https://www.limswiki.org/index.php/Irisnote,_Inc.`
SparkLix Bio IT	SparkLix	Discontinued in Summer 2014. `http://www.limswiki.org/index.php?title=SparkLix_Bio_IT_Corp`
SPL Host	Datacloud	Ceased marketing this product in May 2017. `http://www.datacloudlabs.com/web/index.html`

Symyx	Symyx Notebook	Merged with Elsevier MDL - MDL Notebook.
		`http://en.wikipedia.org/wiki/MDL_Information_Systems`
		Merged with Accelrys, Inc in 2010.
		`http://accelrys.com/about/news-pr/merger-0610.html`

Sysment	Sysment Notebook	Notebook has been removed from main website in 2017. `http://www.sysment.hu/`
Tripos	Benchware Notebook	Merged with other companies to make Certara. `https://www.certara.com/`
VelQuest	SmartLab	Purchased by and absorbed into Accelrys Inc in Jan 2012. `http://www.businesswire.com/news/home/20120103005296/en/Accelrys-Acquires-VelQuest-Corporation-35-Million-Cash#.VWc3RFnBzGc`

A.2 Active Vendors of ELNs

ELN	R & D	QA / QC	Chemistry	Biology	Life Sci	Pharmaceutical	Multi Discipline	All Purpose	Semantic Web
AAC Infotray AG - *Limsophy DoDoc* `http://www.limsophy.com/Webservice/page/M1006/0`	✓								
Abbott Informatics - *STARLIMS* `https://www.abbottinformatics.com/`	✓		✓			✓	✓		
Advanced Chemistry Development - *Electronic Notebook for Academia* `http://www.acdlabs.com/solutions/academia/eln/`			✓						
Agaram Technologies - *LogiLab* `http://www.agaramtech.com/product/logilab-eln.html`		✓							
AgiLab - *ChemELN, ELN BioLab, ELN FormuLab* `http://agilab.fr/`			✓	✓			✓		
AgileBio - *LabCollector* `http://labcollector.com/`				✓	✓	✓	✓		
Agilent Technologies - *OpenLAB ELN* `http://www.agilent.com/`	✓								
Aiderbotics Corporation - *elucidaid* `https://aiderbotics.com/`				✓	✓				
Amphora Research Systems, Inc. - *Patent Safe ELN* `https://www.amphora-research.com/products/patentsafe/`									✓
Arxspan - *ArxLab Electronic Notebook* `http://www.arxspan.com/electronic-lab-notebook.asp`			✓	✓			✓		
Asseco Denmark - *shareSignELN* `http://www.asseco.com/dk/home/sharesigneln/welcome/`								✓	
Asymptotic Automation Solutions LLP - *ELM* `https://www.elabmanager.com/`								✓	
ATGC Labs - *ActiveLN* `http://www.atgclabs.com/`				✓					
Benchling, Inc. - *Benchling* `https://benchling.com/`			✓	✓			✓		
BioChemLab Solutions - *Electronic Lab Notebook* `http://biochemlabsolutions.com/ELN/ELN.html`			✓	✓			✓		
Bio Data - *Labguru* `http://www.labguru.com/`				✓					
Bio-ITech - *eLabJournal* `https://www.bio-itech.nl/elabjournal/`					✓				

ChemBytes - *Espresso ELN* `http://chembytes.com/index.php/espresso/download`			✓				✓		
ChemInnovation Software - *CBIS Notebook* `http://www.cheminnovation.com/brochures/CBIS.pdf`			✓	✓					
CompuDrug International - *Laboratory Manager Plus* `http://www.compudrug.com/laboratory_manager_plus`			✓						
Core Informatics - *Core ELN, Chemistry ELN, Biology ELN* - `http://www.corelims.com/products/core-eln/chemistery-eln/`			✓	✓				✓	
Dassault Systmes - *Accelrys ELN* `http://www.3ds.com/`		✓							
DeltaSoft - *DeltaBook* `http://02d15af.netsolhost.com/docs/eln.pdf`	✓		✓				✓		
DIMA Engineering Pvt. Ltd. - *eLabNotes* `http://www.dimaengineering.com/`			✓			✓			
Dotmatics - *Studies Notebook* `http://www.dotmatics.com/products/studies-notebook/`			✓	✓			✓		
EasyLab Ltd - *EasyLab* `http://easylab.com.tr/`	✓								
Edge Software Consultancy - *BioRails* `http://www.edge-ka.com/products/biorails`				✓		✓			
eNovalys SAS - *ePro* `http://www.enovalys.com/epro/`			✓						
enso Software - *ensochemLab* - `http://www.enso-software.com/website/pages/ensochemlab.de.php`			✓						
Evolvus - *Electronic Lab Notebook* `http://www.evolvus.com/products/itproducts/electroniclabnotebook.html`			✓						
FORMULATOR - *FORMULATOR* `http://www.formulatorus.com/chemists.htm`			✓						
Genohm BVBA - *SLims* `http://www.genohm.com/`								✓	
GoLIMS - *GoLIMS* `http://www.golims.com/`								✓	
HiTec Zang - *eJournal* `http://www.hitec-zang.de/en/laboratory-automation/software/`								✓	
iAdvantage Software - *eStudy* `http://www.iadvantagesoftware.com/wpesoverview.asp`	✓	✓					✓		

iDBS - *E-Workbook* http://www.idbs.com/en/platform-products/e-workbook/								✓	
Instem LSS Limited - *Logbook* http://www.instem-lss.com/	✓								
iVention BV - *iVentionLES* http://www.ivention.nl/	✓				✓				
Kalliste - *Kalliste eSystems* http://www.kallistesystems.com/						✓			
KineMatic - *eNovator ELN* http://www.kinematik.com/	✓	✓						✓	
Labage - *Benchsys* http://www.benchsys.com/	✓								
Lab-Ally LLC - *CERF* http://lab-ally.com/				✓	✓	✓			
Lab-Ally LLC -*RSpace* http://lab-ally.com/				✓	✓	✓			✓
LabArchives - *LabArchives* http://www.labarchives.com/								✓	
LabCollector - *LabCollector ELN* http://www.labcollector.com/								✓	
Labfolder - *labfolder* https://www.labfolder.com/								✓	
Labii, Inc. - *Labii ELN & LIMS* https://www.labii.com/			✓	✓	✓				
LABTrack - *LABTrack* http://www.labtrack.com/	✓	✓						✓	
LabVantage - *eNotebook* http://www.labvantage.com/								✓	
LabWare - *LabWare ELN* http://www.labware.com/en/p/Products/ELN			✓				✓		
Laurus Infosystems - *Chemia* http://www.laurusis.com/	✓								
McNeilCo - *ACAS* http://www.mcneilco.com/acas_docs.html#7				✓					
Mestrelab - *Mbook* http://store.mestrelab.com/store/mbook-cloud.html			✓						
National Institute of Allergy and Infectious Diseases - *LabShare* ttp://www.niaid.nih.gov/			✓		✓				
NoteBookMaker - *NoteBookMaker* http://www.notebookmaker.com/	✓								
Online LIMS - *Online Worksheet* http://www.online-lims.com/								✓	

Open.Co Srl - *Open.Co ELN* `http://www.openco.it/`	✓							✓	
Open Source - *CyNote* `http://cynote.sourceforge.net/`					✓				
Open Source - *eLabFTW* `http://www.elabftw.net/`								✓	
Open Source - *eln* `https://launchpad.net/eln`								✓	
Open Source - *Indigo ELN* `http://ggasoftware.com/opensource/indigo/eln`			✓	✓					
Open Source - *LabTrove* `http://www.labtrove.org/`			✓						
Open Source - *MyLabBook* `http://sourceforge.net/projects/mylabbook/`					✓				✓
Open Source - *Open enventory* `http://www.chemie.uni-kl.de/index.php?id=653`			✓	✓					
Open Source - *OpenBIS* `http://www.cisd.ethz.ch/software/openBIS`					✓				
Open Source - *SciNote* `https://scinote.net/`								✓	
PerkinElmer - *E-Notebook* `http://www.cambridgesoft.com/literature/PDF/ENotebook_2014.pdf`			✓	✓			✓		
Prog4biz Software Solutions Ltd. - *BookitLab* `http://prog4biz.com/`								✓	
Quattro Research - *quattro/Lj* `www.quattro-research.com/quattro-LJ.63.0.html`	✓		✓	✓			✓		
Rescop BV - *RC-ELN* `http://www.rescop.com/`								✓	
RURO - *Sciency ELN* `http://ruro.com/software/sciency-eln/overview`			✓	✓			✓		
Sapio Science - *Exemplar* `https://www.sapiosciences.com/`				✓					
SciCord, LLC - *SciCord* `https://scicord.com/`	✓		✓		✓	✓			
Sciformation - *Sciformation ELN* `http://sciformation.com/sciformation_eln_lj.html?lang=en`			✓	✓					
Scilligence Corp - *Scilligence ELN* `http://www.scilligence.com/web/eln.aspx`			✓	✓					
Seqome Limited - *ELNOME* `http://seqome.com/`				✓					
Shanghai Holo Sci-Infor - *Electronic Lab Notebook* `http://www.holoinfo.com.cn/english/products/elnac.html`			✓	✓			✓		

Shazino - *Hive Bench* `https://www.hivebench.com/`				✓					
Siemens AG - *SIMATIC IT R&D* `http://www.siemens.com/`	✓								
StackWave, LLC - *StackWave ELN* `http://www.stackwave.com/`								✓	
Studylog - *Study Director* `http://www.studylog.com/`				✓					
SunBio - *SunBio ELN* `http://www.sunbioit.com/products/sun-bio-eln/`			✓	✓	✓				
Synbiota - *Synbiota* `https://synbiota.com/#`			✓	✓	✓	✓	✓		
Systat - *SigmaCERF* `http://www.sigmaplot.co.uk/`				✓	✓	✓	✓		
Terrington - *Labsform* `http://www.terringtondm.com/`								✓	
Textco BioSoftware - *Gene Inspector* `http://www.textco.com/`				✓					
Waters - *NuGenesis* `http://www.waters.com/waters/home.htm?locale=en_GB`	✓	✓					✓		

A.3 ELN Licenses / Platforms

	License			Platform					
ELN	Commercial	Open Source	Free Version	Independent	Web Based	Windows	Mac	Unspecified	Other software
AAC Infotray AG - *Limsophy DoDoc* http://www.limsophy.com/Webservice/page/M1006/0	✓					✓			
Abbott Informatics - *STARLIMS* https://www.abbottinformatics.com/	✓			✓	✓				
Advanced Chemistry Development - *Electronic Notebook for Academia* http://www.acdlabs.com/solutions/academia/eln/	✓							✓	
Agaram Technologies - *LogiLab* http://www.agaramtech.com/product/logilab-eln.html	✓					✓			
AgiLab - *ChemELN, ELN BioLab, ELN FormuLab* http://agilab.fr/	✓							✓	
AgileBio - *LabCollector* http://labcollector.com/	✓		✓			✓	✓		✓
Agilent Technologies - *OpenLAB ELN* http://www.agilent.com/	✓		✓			✓			
Aiderbotics Corporation - *elucidaid* https://aiderbotics.com/	✓				✓	✓	✓		
Amphora Research Systems, Inc. - *Patent Safe ELN* https://www.amphora-research.com/products/patentsafe/	✓							✓	
Arxspan - *ArxLab Electronic Notebook* http://www.arxspan.com/electronic-lab-notebook.asp	✓			✓	✓				
Asseco Denmark - *shareSignELN* http://www.asseco.com/dk/home/sharesigneln/welcome/	✓					✓			
Asymptotic Automation Solutions LLP - *ELM* https://www.elabmanager.com/	✓		✓	✓	✓				
ATGC Labs - *ActiveLN* http://www.atgclabs.com/	✓		✓	✓	✓				
Benchling, Inc. - *Benchling* https://benchling.com/	✓		✓	✓	✓				
BioChemLab Solutions - *Electronic Lab Notebook* http://biochemlabsolutions.com/ELN/ELN.html	✓		✓	✓	✓				
Bio Data - *Labguru* http://www.labguru.com/	✓			✓	✓				
Bio-ITech - *eLabJournal* https://www.bio-itech.nl/elabjournal/	✓			✓	✓				

ChemBytes - *Espresso ELN* http://chembytes.com/index.php/espresso/download	✓					✓			✓
ChemInnovation Software - *CBIS Notebook* http://www.cheminnovation.com/brochures/CBIS.pdf	✓					✓	✓		
CompuDrug International - *Laboratory Manager Plus* http://www.compudrug.com/laboratory_manager_plus	✓			✓					
Core Informatics - *Core ELN, Chemistry ELN, Biology ELN* - http://www.corelims.com/products/core-eln/chemistery-eln/	✓			✓	✓				
Dassault Systmes - *Accelrys ELN* http://www.3ds.com/	✓			✓	✓	✓	✓		
DeltaSoft - *DeltaBook* http://02d15af.netsolhost.com/docs/eln.pdf	✓			✓	✓				
DIMA Engineering Pvt. Ltd. - *eLabNotes* http://www.dimaengineering.com/	✓			✓	✓				
Dotmatics - *Studies Notebook* http://www.dotmatics.com/products/studies-notebook/	✓			✓	✓				
EasyLab Ltd - *EasyLab* http://easylab.com.tr/	✓				✓				
Edge Software Consultancy - *BioRails* http://www.edge-ka.com/products/biorails	✓					✓	✓		
eNovalys SAS - *ePro* http://www.enovalys.com/epro/	✓		✓			✓			
enso Software - *ensochemLab* - http://www.enso-software.com/website/pages/ensochemlab.de.php	✓					✓			
Evolvus - *Electronic Lab Notebook* http://www.evolvus.com/products/itproducts/electroniclabnotebook.html	✓							✓	
EZQuant - *EZQuant-ELN* http://www.ezquant.com/en/	✓		✓			✓			
FORMULATOR - *FORMULATOR* http://www.formulatorus.com/chemists.htm	✓					✓			
Genohm BVBA - *SLims* http://www.genohm.com/	✓					✓	✓		✓
GoLIMS - *GoLIMS* http://www.golims.com/	✓			✓	✓				
HiTec Zang - *eJournal* http://www.hitec-zang.de/en/laboratory-automation/software/	✓			✓	✓				

iAdvantage Software - *eStudy* `http://www.iadvantagesoftware.com/wpesoverview.asp`	✓			✓	✓				
iDBS - *E-Workbook* `http://www.idbs.com/en/platform-products/e-workbook/`	✓			✓	✓				
Instem LSS Limited - *Logbook* `http://www.instem-lss.com/`	✓					✓			
iVention BV - *iVentionLES* `http://www.ivention.nl/`	✓			✓	✓				
Kalliste - *Kalliste eSystems* `http://www.kallistesystems.com/`	✓			✓					
KineMatic - *eNovator ELN* `http://www.kinematik.com/`	✓			✓	✓				
Labage - *Benchsys* `http://www.benchsys.com/`	✓							✓	
Lab-Ally LLC - *CERF* `http://lab-ally.com/`	✓					✓	✓		
Lab-Ally LLC -*RSpace* `http://lab-ally.com/`	✓			✓	✓				
LabArchives - *LabArchives* `http://www.labarchives.com/`	✓		✓			✓	✓		
LabCollector - *LabCollector ELN* `http://www.labcollector.com/`	✓			✓	✓				✓
Labfolder - *labfolder* `https://www.labfolder.com/`	✓		✓	✓	✓				
Labii, Inc. - *Labii ELN & LIMS* `https://www.labii.com/`	✓			✓					
LABTrack - *LABTrack* `http://www.labtrack.com/`	✓					✓	✓		
LabVantage - *eNotebook* `http://www.labvantage.com/`	✓			✓	✓				
LabWare - *LabWare ELN* `http://www.labware.com/en/p/Products/ELN`	✓					✓			
Laurus Infosystems - *Chemia* `http://www.laurusis.com/`	✓							✓	
McNeilCo - *ACAS* `http://www.mcneilco.com/acas_docs.html#7`	✓							✓	
Mestrelab - *Mbook* `http://store.mestrelab.com/store/mbook-cloud.html`	✓			✓	✓				
National Institute of Allergy and Infectious Diseases - *LabShare* `ttp://www.niaid.nih.gov/`	✓							✓	
NoteBookMaker - *NoteBookMaker* `http://www.notebookmaker.com/`	✓					✓	✓		

Online LIMS - *Online Worksheet* http://www.online-lims.com/	✓					✓			
Open.Co Srl - *Open.Co ELN* http://www.openco.it/	✓					✓	✓		
Open Source - *CyNote* http://cynote.sourceforge.net/		✓		✓	✓				
Open Source - *eLabFTW* http://www.elabftw.net/		✓		✓	✓				
Open Source - *eln* https://launchpad.net/eln		✓				✓	✓		
Open Source - *Indigo ELN* http://ggasoftware.com/opensource/indigo/eln		✓		✓					✓
Open Source - *LabTrove* http://www.labtrove.org/		✓		✓					
Open Source - *MyLabBook* http://sourceforge.net/projects/mylabbook/		✓							
Open Source - *Open enventory* http://www.chemie.uni-kl.de/index.php?id=653		✓		✓	✓				
Open Source - *OpenBIS* http://www.cisd.ethz.ch/software/openBIS		✓				✓			✓
Open Source - *SciNote* https://scinote.net/		✓		✓	✓				
PerkinElmer - *E-Notebook* http://www.cambridgesoft.com/literature/PDF/ENotebook_2014.pdf	✓							✓	
Prog4biz Software Solutions Ltd. - *BookitLab* http://prog4biz.com/	✓			✓					
Quattro Research - *quattro/Lj* www.quattro-research.com/quattro-LJ.63.0.html	✓					✓			
Rescop BV - *RC-ELN* http://www.rescop.com/	✓			✓					
RURO - *Sciency ELN* http://ruro.com/software/sciency-eln/overview	✓			✓	✓				
Sapio Science - *Exemplar* https://www.sapiosciences.com/	✓			✓	✓				
SciCord, LLC - *SciCord* https://scicord.com/	✓			✓					
Sciformation - *Sciformation ELN* http://sciformation.com/sciformation_eln_lj.html?lang=en	✓			✓	✓				
Scilligence Corp - *Scilligence ELN* http://www.scilligence.com/web/eln.aspx	✓			✓					
Seqome Limited - *ELNOME* http://seqome.com/	✓			✓					

Shanghai Holo Sci-Infor - *Electronic Lab Notebook* `http://www.holoinfo.com.cn/english/products/elnac.html`	✓							✓	
Shazino - *Hive Bench* `https://www.hivebench.com/`			✓	✓	✓				
Siemens AG - *SIMATIC IT R&D* `http://www.siemens.com/`	✓			✓	✓				
StackWave, LLC - *StackWave ELN* `http://www.stackwave.com/`	✓			✓					
Studylog - *Study Director* `http://www.studylog.com/`	✓			✓		✓	✓		
SunBio - *SunBio ELN* `http://www.sunbioit.com/products/sun-bio-eln/`	✓							✓	
Synbiota - *Synbiota* `https://synbiota.com/#`	✓		✓	✓	✓				
Systat - *SigmaCERF* `http://www.sigmaplot.co.uk/`	✓		✓			✓	✓		
Terrington - *Labsform* `http://www.terringtondm.com/`	✓			✓	✓				
Textco BioSoftware - *Gene Inspector* `http://www.textco.com/`	✓		✓			✓	✓		
Waters - *NuGenesis* `http://www.waters.com/waters/home.htm?locale=en_GB`	✓			✓	✓				

Appendix B

Survey

This appendix details the survey that was conducted. It was created on iSurvey[1], and participants were approached by facebook and email; emails were sent out to the following mailing lists:

- https://list.indiana.edu/sympa/info/chminf-l
- molecular-dynamics-news@jiscmail.ac.uk
- spectroscopy-group@jiscmail.ac.uk

B.1 Ethics

This survey was applied for under ethics application ERGO/FPSE/17642 with the following documents:

- Ethics Form
- Participant Information Sheet
- Risk Assessment Form
- Consent Form
- Copy of the Survey

B.2 Supplementary Material

The following supplementary material is available for this study in the pure dataset: `https://doi.org/10.5258/SOTON/D0384`

- Full copy of the survey - ChemistsSoftwareSurvey-FullSurvey.pdf
- List of tools used in the survey - ChemistsSoftwareSurvey-SoftwareTools.xlsx
- Anonymised survey results - ChemistsSoftwareSurvey-Results.xlsx

[1] `https://isurvey.soton.ac.uk/`

B.3 Survey Questions

Section 1 - Demographics

1. Please select all the types of Chemist that you feel apply to you
   - Analytical
   - Environmental
   - Industrial
   - Inorganic
   - Material
   - Organic
   - Physical
2. How many years of chemistry research experience have you had?

Section 2 - Molecular Modelling & Simulation

This section is about your usage of molecular modelling & simulation software

3. Have you ever used molecular modelling & simulation software?

Section 3 - Molecular Modelling & Simulation Software

This section is about your usage of molecular modelling & simulation software programs

4. Please state how often you have used these types of molecular modelling & simulation software (Often, Rarely, or Never)
   - Abalone
   - ACEMD
   - ADUN
   - AMBER
   - Ascalaph Designer
   - Automated Topology Builder
   - Avogadro
   - BALLVIEW
   - Biskit
   - Blaze
   - BOSS
   - CHARMM
   - CHEMKIN
   - Cosmos
   - CP2K
   - Culgi
   - Deneb
   - Desmond
   - Discovery Studio
   - DOCK
   - Extensible Computational Chemistry
   - Environment
   - FastROCS

- Firefly
- FoldX
- Gabedit
- Ghemical
- GOMC
- GPIUTMD
- GROMACS
- GROMOS
- HALMD
- HOOMD
- ICM Suite
- LAMMPS
- Lead Finder
- MacroModel
- Maestro
- MAPS
- Materials Studio
- MedeA Gibbs
- MCCCS Towhee
- MDynaMix MOE
- MOIL
- Molden
- NAB
- NAMD + VMD
- Newton-X
- NWChem
- Octopus
- ORAC
- oxDNA
- Packmol
- pi-qmc
- Prime
- Protein Local Optimization Program
- Q
- Qwalk
- SCIGRESS
- Spartan
- StruMM3D
- TeraChem
- TINKER
- Tremolo-X
- UCSF Chimera
- VASP
- VEGA ZZ

- VLifeMDS
- WHAT IF
- YASARA

Section 4 - Molecular Editor

This section is about your usage of molecular editor software

5. Have you ever used molecular editor software?

Section 5 - Molecular Editor Software

This section is about your usage of molecular editor software programs

6. Please state how often you have used these types of molecular editor software (Often, Rarely, or Never)

   - 3D Molecules Editor Accelrys Draw
   - ACD / ChemSketch Amira
   - ArgusLab
   - Ascalaph Designer Avogadro
   - BALLView
   - Bioeclipse
   - BKChem
   - Chem3D
   - ChemDoodle
   - ChemDraw
   - chemicalize.org
   - ChemJuice
   - ChemTool
   - ChemWindow
   - ChemWriter
   - Chirys Draw / Chirys Sketch CLC Workbench
   - Deneb
   - Elemental
   - ICEDIT
   - JChemPaint
   - JME Molecule Editor
   - JMol
   - JSDraw
   - JSME
   - JSMol
   - KnowItAll
   - Marvin (MarvinSketch, MarvinSpace)
   - MedChem Designer
   - Molecular Editor Software and Image Sharer MolEditor
   - Molinspiration
   - molsKetch
   - MolView

- ODYSSEY
- PubChem
- Rastop
- SketchEl
- Smormo-Ed
- Spartan
- StruMM3D
- Vimol
- XDrawChem

Section 6 - Quantum Chemistry
This section is about your usage of quantum chemistry software

7. Have you ever used quantum chemistry software?

Section 7 - Quantum Chemistry Software
This section is about your usage of quantum chemistry software programs

8. Please state how often you have used these types of quantum chemistry software (Often, Rarely, or Never)
    - ACES ADF
    - AIMAll
    - AMPAC
    - Atomistix ToolKit (ATK)
    - BigDFT
    - CADPAC
    - Car-Parrinello (CPMD)
    - CASINO
    - CASTEP
    - CFOUR
    - COLUMBUS
    - CP2K
    - CRUNCH
    - Crystal
    - DACAPO
    - DALTON
    - DeMon2k
    - DFTB
    - DFT++
    - DIRAC
    - DMol3
    - ELK
    - Empire
    - ErgoSCF
    - ERKALE
    - EXCITING

- FHI-aims
- Firefly
- FreeON
- GAMESS(UK)
- GAMESS(US)
- Gaussian
- Ghemical
- GPAW
- HiLAPW
- HORTON
- ICM Suite
- Jaguar
- JDFTx
- LOWDIN
- MADNESS
- MOLCAS
- MOLPRO
- MOPAC
- MPQC
- NWChem
- Octopus
- ONETEP
- OpenAtom
- OpenMX
- ORCA
- PARSEC
- PSI
- PyQuante
- PySCF
- Q-Chem
- QSite
- Quantemol-N
- Quantum ESPRESSO
- RMG
- SCIGRESS
- Spartan
- Siam Quantum
- SIESTA
- TB-LMTO
- TeraChem
- TURBOMOLE
- VASP
- VB2000
- WIEN2k

- XMVB
- Yambo Code

Section 8 - Organic Synthesis

This section is about your usage of organic synthesis software

9. Have you ever used organic synthesis software?

Section 9 - Organic Synthesis Software

This section is about your usage of organic synthesis software programs

10. Please state how often you have used these types of organic synthesis software (Often, Rarely, or Never)
    - CHIRON
    - ICSynth
    - SYLVIA
    - WODCA

Section 10 - Nanostructure Modelling

This section is about your usage of nanostructure modelling software

11. Have you ever used nanostructure modelling software?

Section 11 - Nanostructure Modelling Software

This section is about your usage of nanostructure modelling software programs

12. Please state how often you have used these types of nanostructure modelling software (Often, Rarely, Never)
    - Ascalaph Designer
    - Atomistix ToolKit (ATK)
    - Atomistix Virtual NanoLab
    - CST STUDIO SUITE
    - CoNTub
    - Deneb
    - Nanohub
    - Ninithi
    - Nanotube Modeller
    - Materials Design MedeA
    - Materials Studio
    - Quantum DotLab
    - SCIGRESS
    - Tubegen
    - Wrapping

Section 12 - Chemical Kinetics & Process Simulator

This section is about your usage of chemical kinetics & process simulator software

13. Have you ever used chemical kinetics & process simulator software?

Section 13 - Chemical Kinetics & Process Simulator Software

This section is about your usage of chemical kinetics & process simulator software programs

14. Please state how often you have used these types of chemical kinetics & process simulator software (Often, Rarely, Never)

    - Cantera
    - ChemCollective
    - Chemical Workbench
    - COSILAB
    - DWSIM
    - Khimera
    - ReactLab KINETICS

Section 14 - Chemical Database & Informatics

This section is about your usage of chemical database & informatics software

15. Have you ever used chemical database & infomatics software?

Section 15 - Chemical Database & Informatics Software

This section is about your usage of chemical database & informatics software programs

16. Please state how often you have used these types of chemical database & informatics software (Often, Rarely, Never)

    - ASU Physical, Chemical and Other Property Data
    - ChemBioFinder
    - ChemExper
    - ChemFinder
    - ChemIDPlus
    - Chemical Development Kit
    - Chem Spider
    - Chemical Workbench
    - Chemical Theasaurus
    - IUPAC-NIST Solubility Database
    - OpenBabel
    - Organic Reactions
    - Organic Syntheses
    - PubChem
    - NIST Chemistry WebBook
    - NIST Chemical Kinetics Database
    - NIST Physical Reference Data
    - Reaxys Database
    - RDKit
    - Sigma-Aldrich Reaction Database
    - ZINC

Section 16 - Chemistry Bibliographic Database

This section is about your usage of chemical bibliographic database software

17. Have you ever used chemistry bibliographic database software?

Section 17 - Chemistry Bibliographic Database Software

This section is about your usage of chemical bibliographic database software programs

18. Please state how often you have used these types of chemistry bibliographic database software (Often, Rarely, Never)
    - Analytical Abstracts Database
    - OJOSE: Online JOurnal Search Engine
    - Synthesis Reviews

Section 18 - Computer Based Chemical Terminology (Semantic Web)

This section is about your usage of computer based chemical terminology (semantic web) software

19. Have you ever used any computer based chemical terminology (semantic web) software?

Section 19 - Computer Based Chemical Terminology (Semantic Web) Software

This section is about your usage of computer based chemical terminology (semantic web) software programs

20. Please state how often you have used these types of computer based chemical terminology (semantic web) software (Often, Rarely, Never)
    - Avogadro
    - ChEBI - The Database & Ontology of Chemical Entities of Biological Interest
    - Chemical Tagger
    - CHEMINF - Chemical Information Ontology
    - CMO - Chemical Methods Ontology
    - MOP - Molecular Processes Ontology
    - NanoParticle Ontology
    - RXNO - Name Reaction Ontology

Section 20 - Other

This section is about your usage of other chemistry software

21. Have you ever used any other types of chemistry software?

Section 21 - Other Software

This section is about your usage of other chemistry software programs

22. Please state how often you have used these other types of chemistry software (Often, Rarely, Never)
    - A More Accurate Fourier Transform
    - APBS
    - Aqion
    - CIF2Cell
    - DISCUS
    - GaussSum
    - GenX
    - Insensitive
    - OpenChrom
    - PyMca
    - RubyChem

- spgLib
- ToxTree

22 - Future Tools

23. What tool would you most like to be created?

Appendix C

Focus Groups

This appendix details the focus groups that were conducted. A total of 4 biologists, 6 physicists and 14 chemists were involved in these focus groups.

- Chemistry Focus Group 1: 8 Participants: G, H, I, J, K, L, M &N
- Chemistry Focus Group 2: 6 Participants: S, T, U, V, W & X
- Physics Focus Group: 6 Participants: A, B, C, D, E, & F
- Biology Focus Group: 4 Participants: O, P, Q & R

C.1 Ethics

This study was applied for under ethics application ERGO/FPSE/18246 with the following documents:

- Ethics Form
- Participant Information Sheet
- Data Protection Plan
- Risk Assessment Form
- Consent Form

C.2 Supplementary Material

The following supplementary material is available for this study in the pure dataset: `https://doi.org/10.5258/SOTON/D0384`

- Anonymised Focus Group Transcriptions - FocusGroupTranscriptsAnonymised.pdf

C.3 Focus Group Questions

1. What method do you use to record your notes?

2. For each of these following different types of work, what pieces of information do you currently record, and how do you record it (e.g notes, mind maps, graphs, pictures, photos, diagrams, tables etc)
    - Doing an experiment in the lab
    - Doing an experiment outside of the lab
    - Looking at literature
    - Thinking about your work
    - Performing calculations to support your research
    - Writing up your work
3. When taking your notes, how do you organise them? (indexing, creating sections etc)
4. Do you use any technology to aid with your note recording? (instruments: tablets, phones, cameras, recording equipment). *Also prompt for inadvertent use of technology such as emailing yourself?*
5. How do you link any digital resources or notes to paper based notes?
6. Where is your data / research output stored?
7. Are you concerned about IP?
    - Do your records or notes need to be kept secure?
    - Are there limits on who you can share your data with?
    - Does your data need to be kept for a specific period of time?
    - Does your data require any 3rd party sign off?
8. Who do you collaborate with for work, and who do you share your work with?
    - Do you share your work for feedback?
    - Is sharing your work useful?
    - Are there people you need to be able to share your work with?
    - Before you share your work, do you write up your notes or change the format first?
9. Do you use reference management software, if so what?
10. Imagine you're trying to locate a piece of work or some notes from 6 months ago?
    - How would you locate these notes?
    - How would you locate your data?
11. Imagine that there is a fire in your lab and all of your paper notebooks are destroyed?
    - How much of your work would be lost?
    - How could you go about recovering this work?
12. If you fell under a bus tomorrow and were indisposed for a while, how would your supervisor/industry sponsors/colleagues access your work?
13. Where are all of your notes backed up, Electronic and Paper?
14. Have you used ELNs before?
    - What did you like and didn't you like?
    - If you did use one, and you stopped using one, why did you stop?
15. What would you expect that an ELN would be able to do for you?
16. How could an ELN make recording your work better?
17. What equipment are you allowed to take into the lab?
18. Do you have any further comments on ELNs and Notetaking in the lab in general?

Appendix D

Participant Observations

This appendix details the participant observations that were conducted. Four different labs were observed.

- Crysallography Lab - Participants W, AF & AG
- Molecular Chemistry Lab - Participants S, T & AH
- Organic Chemistry Lab - Participants AI, Y & Z
- Inorganic Chemistry Lab - Participants AA, AB, AC, AD & AE

D.1 Ethics

This study was applied for under ethics application ERGO/FPSE/18448 with the following documents:

- Ethics Form
- Participant Information Sheet
- Data Protection Plan
- Risk Assessment Form
- Consent Form

D.2 Supplementary Material

The following supplementary material is available for this study in the pure dataset: `https://doi.org/10.5258/SOTON/D0384`

- Anonymised write ups of the observations - ObservationsWriteupsAnonymised.pdf

Appendix E

Cloud Notebook Comparison

Functional Requirements / Cloud Notebook Platform	Google Drive	OneNote	Evernote
File Uploads / Downloads	✓	✓	✓
Automatic syncing/updating between devices	✓	✓	✓
Integrate with common data formats	✓	✓	✓
Table of Contents/Index/Highlights	✓	✓	✓
Spell Checker	✓	✓	✓
Keyword Search	✓	✓	✓
Diagrams	✓	✓	✗
Drawing	✓	✓	✗
Text Recongition	✓	✓	✓
Reference Manager links	✓	✓	✗
Recent Activity Feed	✓	✓	✗
TODO Lists	✓	✓	✓
Postit Notes	✗	✓	✗
Print Documents	✓	✓	✓
Voice Capture	✓	✗	✗
Moderation of Comments	✓	✗	✗
Different built in languages	✓	✓	✓
Digital Pen Input	✓	✓	✓
Timelines	✓	✗	✗
Bulletin/Message Boards	✗	✗	✗
Page Statistics	✓	✓	✗
Default Values	✗	✗	✗
Generate Report Burtton	✗	✗	✗
Notifications about approval/Signoff	✓	✗	✗

Table E.1: Comparison of Cloud Notebook Platforms against notebook functional requirements

Appendix F

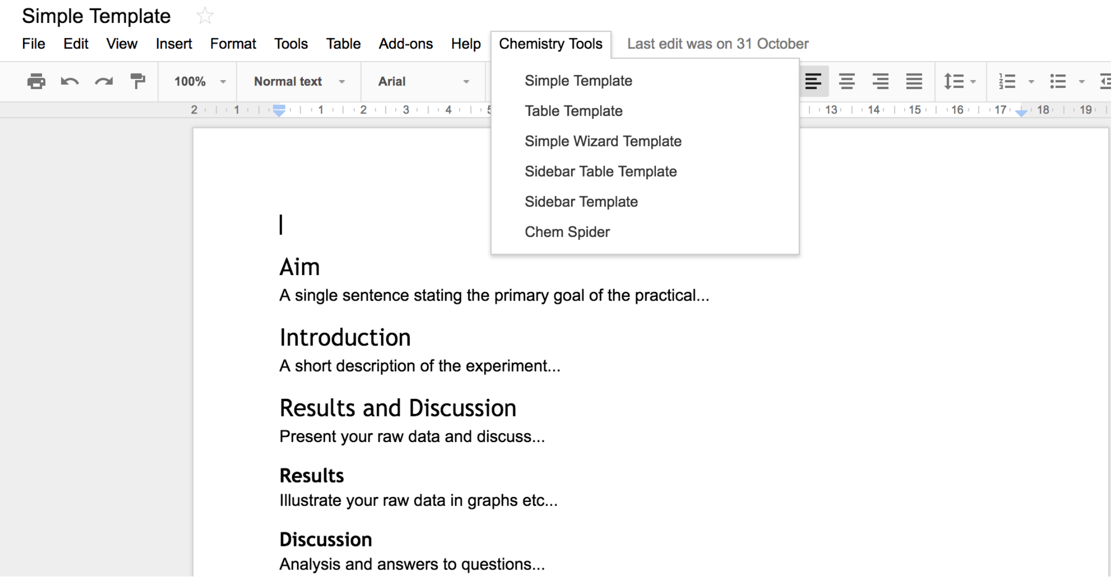
Semanti-Cat

This appendix details the coding repository of the prototype Semanti-Cat, which was created as part of this thesis. The full code for this project can be found at `https://github.com/samikanza/semanti-cat.git`

F.1 Experimental Coding

The basic features of Google App Scripts were tested by creating some custom menus, custom dialog boxes and custom sidebars and integrating a 3rd party application.

F.1.1 Custom Menus



Aim

A single sentence stating the primary goal of the practical...

Introduction

A short description of the experiment...

Results and Discussion

Present your raw data and discuss...

Results

Illustrate your raw data in graphs etc...

Discussion

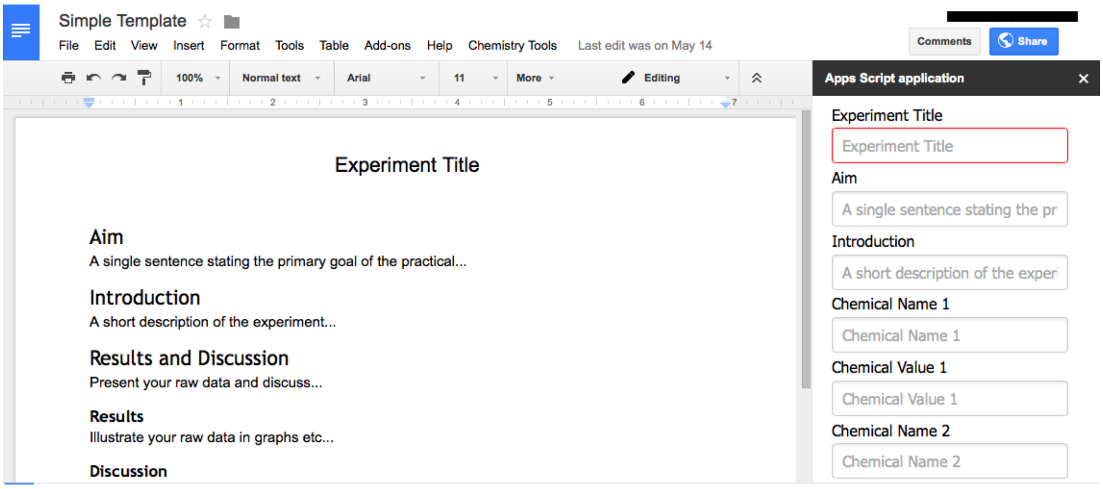
Analysis and answers to questions...

Figure F.1: Custom Menus

Figure F.1 shows the creation of a custom menu, any number of custom menu bars and menu items can be created so it would be possible to create different sets of menus for different add-ons, with different child tasks. This would help organise the available add-ons and give the user useful information about what tasks they wanted to perform. Additionally, this is a non-intrusive way of adding in extra functionality, that the user can make use of if they choose, but it doesn't impact on using the document in the normal way either.

F.1.2 Custom Sidebars & Dialogs

Figure F.2 shows the creation of a custom sidebar. A sidebar that can be used to update different elements in the main document (which could be used to facilitate certain templates or ensure that users input specific pieces of information). The custom sidebars can serve HTML so they can be customised to whatever functionality the developer wished. The main advantage to the sidebars is that they can exist at the side of the document and can be used in conjunction with the document as opposed to overlaying them (as is the case with the custom dialog boxes, see Figure F.5) and requiring closure before the document can be used. However, the disadvantage is that they have a restricted width which means that they may not be suitable for certain tasks (e.g. an add on for drawing chemical structures would probably need more space). However, this could potentially be used to store a set of icons to link to different add-ons as an alternative to using menu items, or to facilitate template entries. Custom Dialogs work on the same principle but instead of existing at the side they produce a new window over the document (see Figure F.5, and have contrasting advantages and disadvantages to the sidebars. They are larger, so there is a wider scope for what can be put inside them (e.g. an add on for chemical structure diagrams) but they are more intrusive than the sidebars.



[redacted]

Experiment Title

Aim

A single sentence stating the primary goal of the practical...

Introduction

A short description of the experiment...

Results and Discussion

Present your raw data and discuss...

Results

Illustrate your raw data in graphs etc...

Discussion

Experiment Title: Experiment Title

Aim: A single sentence stating the pr

Introduction: A short description of the exper

Chemical Name 1: Chemical Name 1

Chemical Value 1: Chemical Value 1

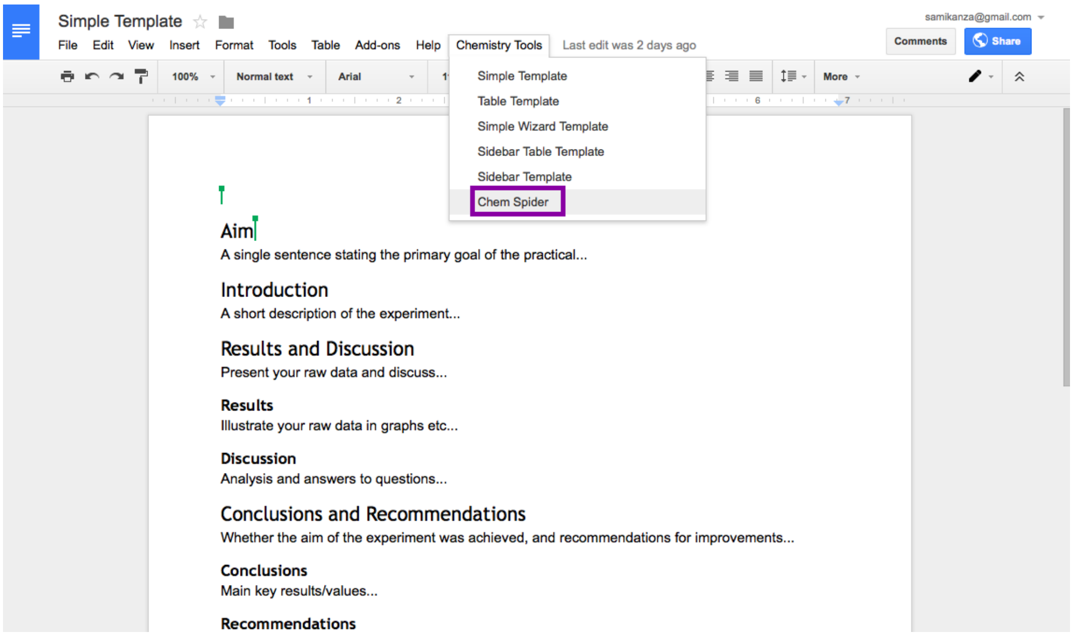
Chemical Name 2: Chemical Name 2

FIGURE F.2: Custom Sidebars

F.1.3 3rd Party Software Integration

ChemSpider was identified in the survey as one of the most popular pieces of database and informatics software, and given it's free usage and API for developers it was chosen to trial as a

3rd party integration. The following figures (F.3, F.4, F.5) show the preliminary integration of ChemSpider[1].



Aim

A single sentence stating the primary goal of the practical...

Introduction

A short description of the experiment...

Results and Discussion

Present your raw data and discuss...

Results

Illustrate your raw data in graphs etc...

Discussion

Analysis and answers to questions...

Conclusions and Recommendations

Whether the aim of the experiment was achieved, and recommendations for improvements...

Conclusions

Main key results/values...

Recommendations

FIGURE F.3: Trial Integration of Chem Spider as a 3rd Party Application

Figure F.3 shows a custom menu with a link to the ChemSpider add-on. This menu item has been programmed such that clicking it will open up a custom dialog with the next part of the ChemSpider integration in it.

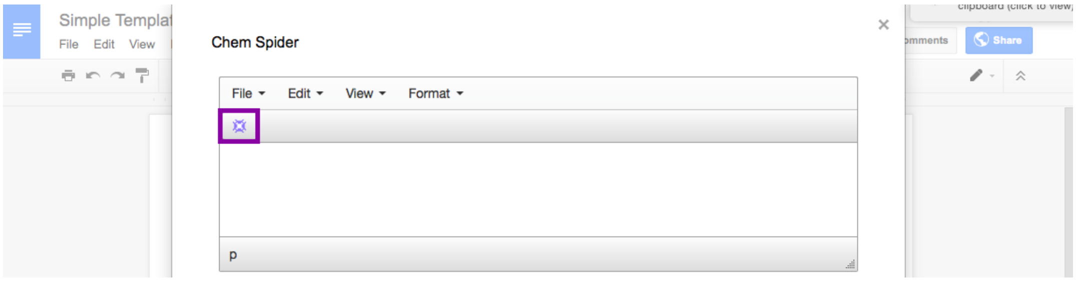


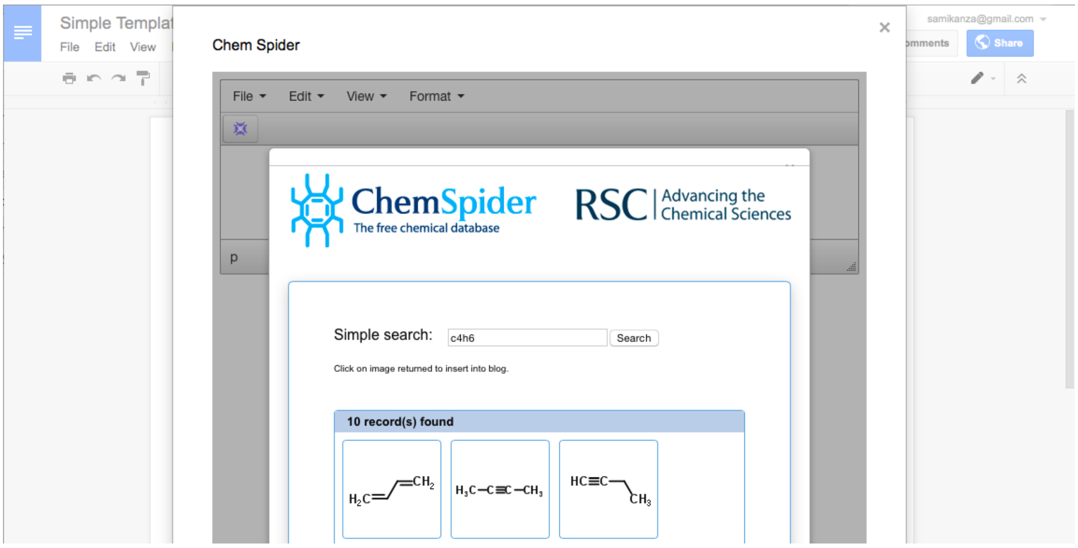
FIGURE F.4: Trial Integration of Chem Spider as a 3rd Party Application

Figure F.4 shows a tinyMCE editor[2] with the ChemSpider plugin as that is currently what it is written for. The work leading up to this integration was to create a standalone page with the tinyMCE editor and get the ChemSpider plugin to work successfully with it, which required making sure all of the appropriate libraries were downloaded and that the folder structure was set up correctly for the scripts to access the correct paths. Once this was achieved, the code was then moved into the Google App Script project bound to this document. If this was to be developed further to be part of the domain layer of the conceptualised ELN environment from this thesis, the code would be decoupled to enable ChemSpider to be directly linked to the Google Doc rather than having to go through the tinyMCE editor first.

[1] http://www.chemspider.com/

[2] http://www.tinymce.com/

The fact that a tinyMCE editor can be integrated in however could prove useful. Preliminary investigations into the document structure of Google Docs suggests that a certain amount of document information can be stored behind the scenes. However, if a particular add-on, for example a Semantic based one, required a higher level of data manipulation or the ability to store additional data then this could be one way of achieving that.



Simple search: c4h6

Click on image returned to insert into blog.

10 record(s) found

FIGURE F.5: Trial Integration of Chem Spider as a 3rd Party Application

Figure F.5 shows the full ChemSpider plugin. This links to the ChemSpider API and facilitates a search of chemical elements. Once the user has chosen the appropriate image of the element then it can be inserted into the document. There is also the potential for combining the service with semantic web technologies as the images could have semantic metadata against them. These methods could be used to encorporate any other types of services that are composed of html and javascript. Additionally this could be used to link to Semantic Web services using any semantic web javascript libraries.

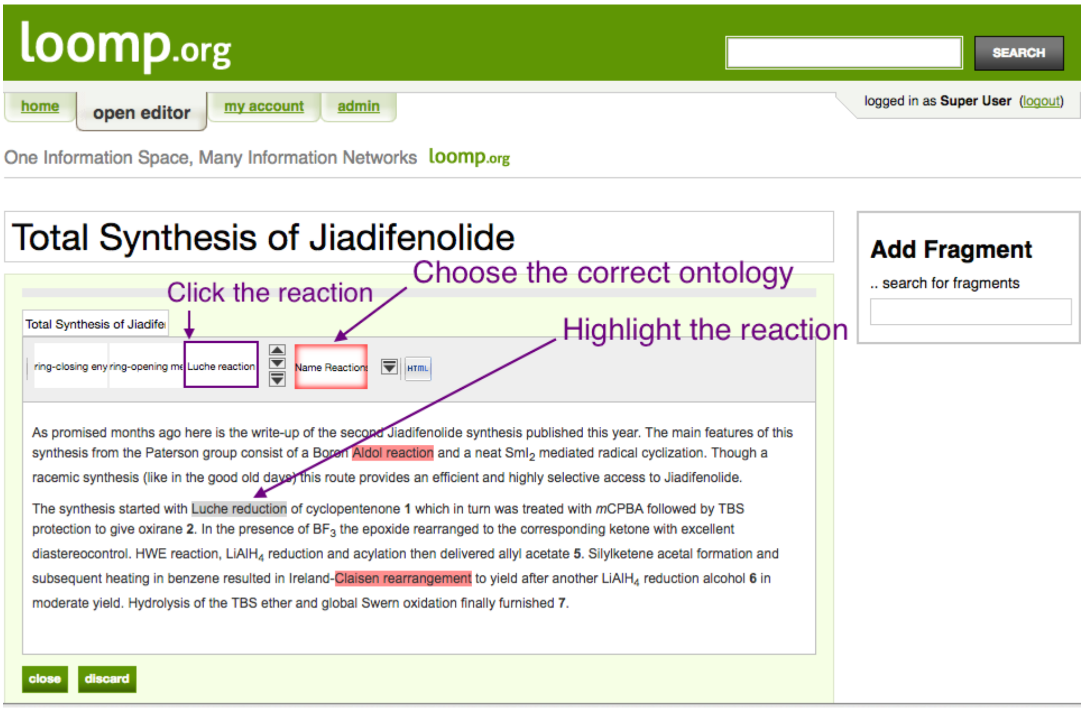
F.1.4 Difficulties

There main difficulty faced when programming with Google App Scripts. The first is its security policy. Whilst it is possible to create HTML files to add to a scripting project, you cannot make JavaScript or CSS files. This leads to potential difficulties for anything that involves large degrees of either as whilst it is possible to embed both JS and CSS in a HTML page it can become very messy and unreadable. Additionally, attempting to host these scripts elsewhere (e.g on a University web space) caused cross side scripting errors. However, A solution to this is to use Google Drive to publish your JS and CSS to the Web and hot-link to them. This worked around the problem as there were no cross-side scripting errors and it still loaded quickly.

F.2 Semantic Mark-up Investigations

A tool called LOOMP[3] that was developed at the University of Southampton has been experimented with to investigate how to incorporate semantic web technologies into the ELN environment. LOOMP currently facilitates marking up fragments of data using ontologies, in this instance geographical ones. This investigation involved taking three ontologies from the Royal Society of Chemistry, the Named Reactions Ontology [4], the Chemical Methods Ontology [5] and the Molecular Process Ontology [6]; and then adding them to this tool. The following example illustrates marking up some chemical data using these ontologies.

Step 1: In the fragment, select the words to mark-up, in this case 'luche reaction' and then find the correct ontology (Named Reactions) and locate the matching reaction in the toolbar.



loomp.org
SEARCH
home
open editor
my account
admin
logged in as Super User (logout)
One Information Space, Many Information Networks loomp.org

Total Synthesis of Jiadifenolide

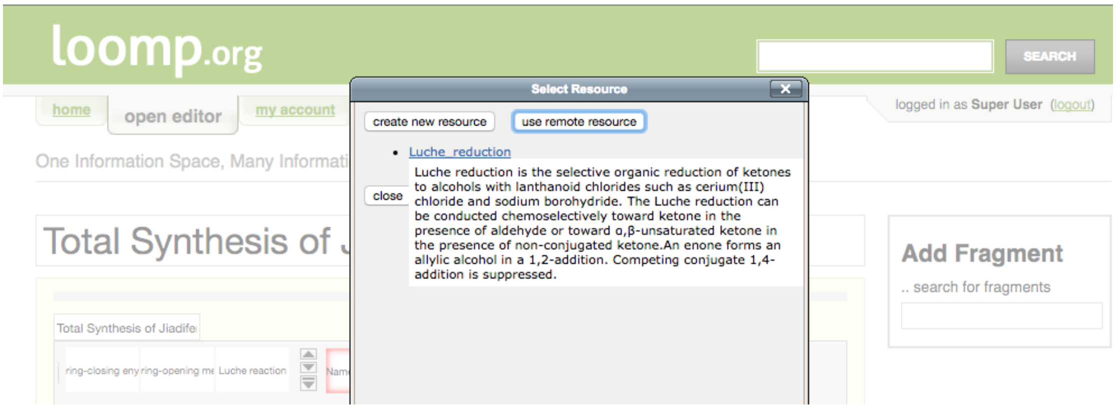
As promised months ago here is the write-up of the second Jiadifenolide synthesis published this year. The main features of this synthesis from the Paterson group consist of a Boron Aldol reaction and a neat SmI_2 mediated radical cyclization. Though a racemic synthesis (like in the good old days) this route provides an efficient and highly selective access to Jiadifenolide.

The synthesis started with Luche reduction of cyclopentenone **1** which in turn was treated with *m*CPBA followed by TBS protection to give oxirane **2**. In the presence of BF_3 the epoxide rearranged to the corresponding ketone with excellent diastereocontrol. HWE reaction, $LiAlH_4$ reduction and acylation then delivered allyl acetate **5**. Silylketene acetal formation and subsequent heating in benzene resulted in Ireland-Claisen rearrangement to yield after another $LiAlH_4$ reduction alcohol **6** in moderate yield. Hydrolysis of the TBS ether and global Swern oxidation finally furnished **7**.

Figure F.6: Loomp Step 1

[3] https://github.com/ag-csw/loomp
[4] https://github.com/rsc-ontologies/rxno
[5] https://github.com/rsc-ontologies/rsc-cmo
[6] https://github.com/rsc-ontologies/rxno/blob/master/mop.obo

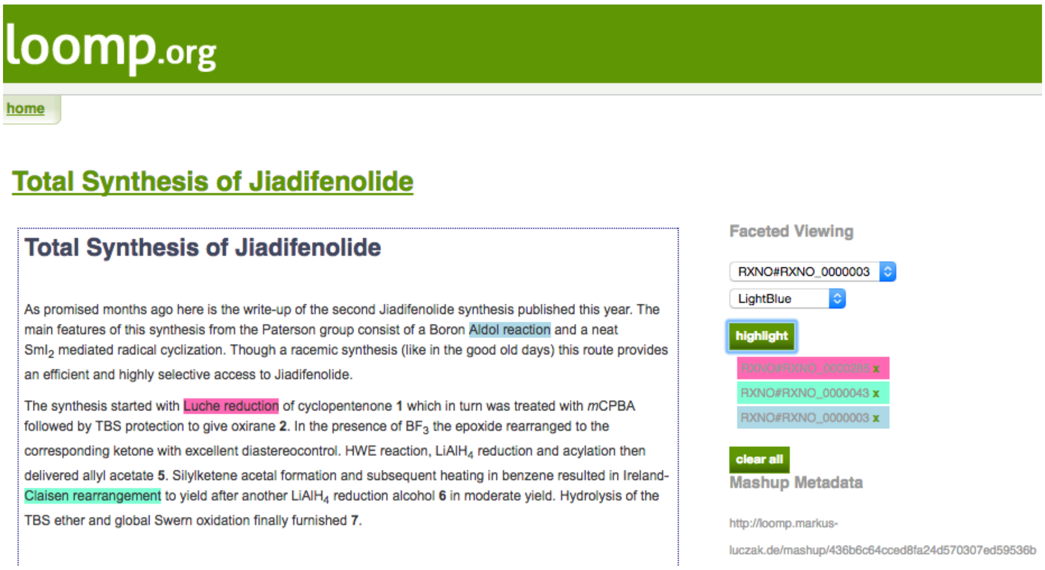
Step 2: Click on the reaction, and then select use remote resource button and it will bring up the information about the reaction to check that it's the right data that you want to associate with this resource.



Luche reduction is the selective organic reduction of ketones to alcohols with lanthanoid chlorides such as cerium(III) chloride and sodium borohydride. The Luche reduction can be conducted chemoselectively toward ketone in the presence of aldehyde or toward α,β-unsaturated ketone in the presence of non-conjugated ketone.An enone forms an allylic alcohol in a 1,2-addition. Competing conjugate 1,4-addition is suppressed.

FIGURE F.7: Loomp Step 2

Step 3: Then you can view the resource and highlight the different pieces of marked up data.



Total Synthesis of Jiadifenolide

As promised months ago here is the write-up of the second Jiadifenolide synthesis published this year. The main features of this synthesis from the Paterson group consist of a Boron Aldol reaction and a neat SmI_2 mediated radical cyclization. Though a racemic synthesis (like in the good old days) this route provides an efficient and highly selective access to Jiadifenolide.

The synthesis started with Luche reduction of cyclopentenone **1** which in turn was treated with *m*CPBA followed by TBS protection to give oxirane **2**. In the presence of BF_3 the epoxide rearranged to the corresponding ketone with excellent diastereocontrol. HWE reaction, $LiAlH_4$ reduction and acylation then delivered allyl acetate **5**. Silylketene acetal formation and subsequent heating in benzene resulted in Ireland-Claisen rearrangement to yield after another $LiAlH_4$ reduction alcohol **6** in moderate yield. Hydrolysis of the TBS ether and global Swern oxidation finally furnished **7**.

http://loomp.markus-luczak.de/mashup/436b6c64cced8fa24d570307ed59536b

FIGURE F.8: Loomp Step 3

Appendix G

Ontologies

G.1 Generic Ontologies

FOAF - Friend of a Friend (Describes people related terms) http://www.foaf-project.org/

DBpedia - (Takes stuff from Wikipedia, NB: has the concept of an ORCID ID) - http://wiki.dbpedia.org/

SKOS Ontology - used to distinguish links between ChEBI and ChEMBL and DrugBank as part of the Open PHACTS project - `https://www.w3.org/TR/2008/WD-skos-reference-20080829/skos.html`

VOID - Vocabulary of Interlinked Datasets - used to specify what predicates are used and what sorts of subjects and objects are interlinked - `http://semanticweb.org/wiki/VoID.html`

G.2 Chemistry Ontologies

RXNO - Name Reaction Ontology (Connected organic name reactions to their roles in an organic synthesis) - `http://www.rsc.org/ontologies/RXNO/index.asp`

CMO- Chemical Methods Ontology (Describes methods, instruments and some material artefacts used in chemical experiments) - `http://www.rsc.org/ontologies/CMO/index.asp`

MOP - Molecular Processes Ontology (Describes chemical processes that take place at the molecular level, e.g methylations and electron transfer) - `http://www.rsc.org/ontologies/MOP/index.asp`

ChEBI - The database and ontology of Chemical Entities of Biological Interest (Dictionary of molecular entities focused on small chemical compounds) - `http://www.ebi.ac.uk/chebi/`

CHEMINF Ontology - Includes terms for the descriptors commonly used in cheminformatics software applications and the algorithms which generate them - `https://bioportal.bioontology.org/ontologies/CHEMINF`

G.3 Biology Ontologies

Plant Ontology - This ontology links together plant related terms - `http://www.obofoundry.org/ontology/po.html`

Cell Ontology - This ontology contains a structured vocabulary for animal cell types - `http://obofoundry.org/ontology/cl.html`

NanoParticle Ontology - An ontology that represents the basic knowledge of physical, chemical and functional characteristics of nanotechnology as used in cancer diagnosis and therapy - `https://bioportal.bioontology.org/ontologies/NPO`

Gene Ontology - The framework for the model of biology. The GO defines concepts/classes used to describe gene function, and relationships between these concepts. - `http://geneontology.org/`

G.4 Physics Ontologies

Astrophysics Ontology - Ontology to cover classes and properties typically used by astronomers - `https://www.astro.umd.edu/~eshaya/astro-onto/ontologies/physics.html`

Appendix H

Software Evaluation Focus Groups

This appendix details the software evaluation focus groups that took place. A total of 5 biologists, 4 physicists and 6 chemists were involved in these focus groups.

- Trial Focus Group: 3 Participants A, B & R
- Chemistry Focus Group: 6 Participants L, J, S, AP, AJ & AK
- Physics Focus Group: 4 Participants A, B, AL & AM
- Biology Focus Group: 5 Participants Q, R, AN, AO & AQ

H.1 Ethics

This study was applied for under ethics application ERGO/FPSE/30211 with the following documents:

- Ethics Form
- Participant Information Sheet
- Data Protection Plan
- Risk Assessment Form
- Consent Form

H.2 Supplementary Material

The following supplementary material is available for this study in the pure dataset: `https://doi.org/10.5258/SOTON/D0384`

- Anonymised Focus Group Transcriptions - SoftwareEvaluationTranscriptsAnonymised.pdf

H.3 Focus Group Script & Questions

"I have been looking at electronic lab notebooks for my PhD thesis, I did a bunch of initial studies, I looked at what was out there in the market, I observed people in the chemistry labs and I did some initial focus groups with physicists chemists and biologists, my interim conclusions were that there's a lot of good reasons to digitise your lab data, there's still a lot of reasons why we should be looking to do that, and that people do actually go and use quite a lot of software for what they're doing. But ultimately there's still this kind of barrier to actually using electronic lab notebooks in the lab, and there's quite a lot of barriers trying to replace paper, so I'm not really looking to try and create a system to replace paper because I don't think that's particularly feasible right now. But what I am looking at is trying to improve the current software offerings of the software that you would use if you were actually writing things up on a computer, and hopefully by improving those we might encourage people to digitise their work further. So, my initial thoughts were when you look at everything, ELNs aren't very popular even though there are loads of them, but electronic notebook software like Google Drive, OneNote, Evernote etc is a lot more popular, and if you break down the features that you would need from an ELN and a EN there's actually a hell of a lot of overlap. For an ELN you need everything you'd have in something like Google Drive and OneNote and then you'd have a domain layer with domain applicable tools, and then a further semantic layer on top that handles metadata and tagging."

The layer diagram illustrated in Figure 4.3 is shown here.

"This is a diagram of the breakdown of the three different layers, being everything here is generic notebook functionality, this would be domain based stuff, and then at the top is the semantic layer which looks at tagging your notes, doing a better search, linking things together, recognising chemicals etc. The bit I've focused on making was the semantic layer. It's all well and good for people to say that they like their documents to be tagged and made better to search, and to have more information about them, but no-ones actually given a great deal of information about what they actually want and how they want it to be achieved. This is a proof of concept system that I've designed. I'm going to show you this in two parts. There is a backend part of the system that shows a representation of what can be done. It shows what tags have been pulled out for each document, which bits of the documents have been marked up, what ontology terms have been recognised, and where different chemicals have been identified. The front end of this system would be in Google Docs for this proof of concept idea, and so you will be shown a mockup of how this functionality could be implemented on the front end. This focus group is to look at both aspects. To investigate how well have the documents been tagged, how things have been marked up etc and also what you guys think of how the front end looks and how you think that would work, and to talk a bit about ELNs.".

Software Evaluation

1. General System
   - (a) What type of document have you brought and how typical a representation of your working documents is it?
   - (b) How much of your documents are usually made up of words / pictures / diagrams?
   - (c) What formats are your documents usually in when you create them?

"Moving onto the tags. These are the tags that have come out of different systems (NB: Each participant was given a copy of their personal documents and the tags as well as being able to

see it on the system). OpenCalais looks at wikipedia articles that have been tagged and looks at what content has been tagged to try and infer how other articles with similar content should be tagged. These are categories of your documents rather than things that have been pulled out from the text. ChemTagger and Gate are the automatic chemical recognition pieces of software so they have pulled out what they think are chemicals in your pieces of work. ChemTagger also identifies actions (gives example showing the marked up actions in backend). As you can see Gate typically seems to pick up less than ChemTagger. For the ontologies, ontologies are basically hierarchies of relationships and concepts that we use in the semantic web so I've currently just got 6 ontologies in here, some chemistry physics and biology ones. This piece will pull out the matching ontology terms. Your documents have been run through all of these systems and currently pulled out everything it can find, and part of this software evaluation is looking to refine this process to figure out what things we should actually be pulling out. So on that note if you all have a brief look at the tags for your categories. "

1. Tags
   (a) Which of the tagging services do you think has the most appropriate tags for your work?
   (b) From the perspective of automatically recognising the chemicals you have used, do you think ChemTagger / Gate have done a good job?
   (c) Are there any of the tagging services you don't think look useful?
   (d) How would you have tagged your document if you'd done it yourself?
   (e) Are there any obvious tags missing?
   (f) Would you make use of an option to add/remove your own tags?
3. Search
   (a) Does the search work as you'd expect?
   (b) When searching for a word, would you expect it to prioritise results by:
      i. Existence in the title
      ii. Tag Weighting
      iii. Term Frequency
   (c) Is being able to search just by title or text useful?
   (d) Would you expect to see any other advanced type of search?
4. Markup
   (a) What do you think of the markup?
   (b) Are the tooltips useful? - most specifically to see the ontology descriptions?
   (c) Would you make use of the ability to edit the tag / chemical descriptions yourself?
   (d) Do you have any suggestions for a better way to markup the documents?
5. Design of the System
   (a) What do you think of the current design of the system?

ELN Behaviour

6. Previous studies have suggested that scientists have a greater tendency to use software for writing up papers/reports/thesis rather than during their lab experiments to take notes, what do you think of this?
7. Would you be more inclined to consider using an ELN if:
   (a) it was aimed at that stage of the lab process (e.g for writing up reports / data analysis etc).

   (b) It semantically tagged your work?
   (c) Facilitated a more accurate search across your work?

8. How you think that semantically tagging your documents would have an impact on the efficiency of your work?
9. Would the following features encourage you to further digitise your work or add tags of your own?
   (a) Semantically tagged documents
   (b) Stronger search
10. Do you have any further comments?

Appendix I

Document Descriptions, Lengths & Tags

Participant / Document Information	Discipline	Document Type	# Pages	# Open Calais Tags	# Ontology Tags	# ChemicalTagger Tags	# Gate Tags
Participant L	Chemistry	Thesis Section	4	12	17	16	4
Participant J	Chemistry	Blog Post	2	6	2	15	0
Participant S	Chemistry	ESI Document	12	12	75	92	3
Participant AP	Chemistry	Thesis Section	4	12	61	62	4
Participant AJ	Chemistry	Handwritten lab book page	1	3	5	10	0
Participant AK	Chemistry	Experimental Writeup	14	7	32	26	5
Participant A	Physics	Research Paper	8	12	79	92	2
Participant B	Physics	Sample Report	4	8	13	8	0
Participant AL	Physics	Handwritten lab book page	1	0	1	3	0
Participant AM	Physics	Handwritten lab book page	1	8	4	0	0
Participant Q	Biology	Thesis Section	1	10	18	8	0
Participant R	Biology	Thesis Section	3	12	25	12	0
Participant AN	Biology	Experimental Writeup	12	7	34	60	0
Participant AO	Biology	Literature Report	10	12	38	58	1
Participant AQ	Biology	Experimental Writeup	7	12	19	20	1

Table I.1: Comparison of the document types, lengths, and how many of each type of tag was applied to it by Semanti-Cat

Bibliography

Abras, C., Maloney-Krichmar, D. and Preece, J. (2004), 'User-Centered Design', *Bainbridge, W. Encyclopedia of Human-Computer Interaction. Thousand Oaks: Sage Publications* **37**(4), 445–456.

Arnstein, L., Hung, C.-Y., Franza, R., Zhou, Q. H., Borriello, G., Consolvo, S. and Su, J. (2002), 'Labscape: A Smart Environment for the Cell Biology Laboratory', *Pervasive Computing, IEEE* **1**(3), 13–21.
URL: *http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1037717*

Atrium Research (2018), Electronic Laboratory Notebook. [Online: Accessed 16-Jan-2018].
URL: *http://www.atriumresearch.com/eln.html*

Babbie, E. (1973), 'Survey Research Methods', *C. Blackstrom and G. Hursh-Cesar (1981) Survey Research.(second edition). New York: Wiley* pp. 297–298.

Badiola, K. A., Bird, C., Brocklesby, W. S., Casson, J., Chapman, R. T., Coles, S. J., Cronshaw, J. R., Fisher, A., Frey, J. G., Gloria, D., Grossel, M. C., Hibbert, D. B., Knight, N., Mapp, L. K., Marazzi, L., Matthews, B., Milsted, A., Minns, R. S., Mueller, K. T., Murphy, K., Parkinson, T., Quinnell, R., Robinson, J. S., Robertson, M. N., Robins, M., Springate, E., Tizzard, G., Todd, M. H., Williamson, A. E., Willoughby, C., Yang, E. and Ylioja, P. M. (2014), 'Experiences with a Researcher-Centric ELN', *Chemical Science* .
URL: *http://xlink.rsc.org/?DOI=C4SC02128B*

Bernard, H. R. and Gravlee, C. C. (2014), *Handbook of Methods in Cultural Anthropology*, Rowman & Littlefield.

Berners-Lee, T., Hall, W., Hendler, J., James, A., O'Hara, K., Shadbolt, N. and Weitzner, D. J. (2006), 'A Framework for Web Science', *Foundations and Trends in Web Science* **1**(1), 1–130.

Berners-Lee, T., Hendler, J., Lassila, O. et al. (2001), 'The Semantic Web', *Scientific American* **284**(5), 28–37.

Bird, C. L., Willoughby, C. and Frey, J. G. (2013), 'Laboratory Notebooks in the Digital Era: the Role of ELNs in Record Keeping for Chemistry and Other Sciences', *Chemical Society Reviews* **42**(20), 8157.
URL: *http://xlink.rsc.org/?DOI=c3cs60122f*

Borkum, M., Lagoze, C., Frey, J. and Coles, S. (2010), A Semantic eScience Platform for Chemistry, *in* 'e-Science (e-Science), 2010 IEEE Sixth International Conference on', pp. 316–323.

Britten, N. (1995), 'Qualitative Research: Qualitative Interviews in Medical Research', *Bmj* **311**(6999), 251–253.

Britten, N. (2007), 'Qualitative Interviews', *Qualitative Research in Health Care, Third Edition* pp. 12–20.

Chen, S. Y., Fan, J.-P. and Macredie, R. D. (2006), 'Navigation in Hypermedia Learning Systems: Experts vs. Novices', *Computers in Human Behavior* **22**(2), 251 – 266.
URL: *http://www.sciencedirect.com/science/article/pii/S0747563204001153*

Clark, A. (2014), 'Putting Together the Pieces: Building a Reaction-Centric Electronic Lab Notebook for Mobile Devices', *Journal of Cheminformatics* **6**(Suppl 1).

Clark, V. P. and Creswell, J. W. (2011), 'Designing and Conducting Mixed Methods Research', *vol* **3**, 93–94.

Coles, S. J., Frey, J. G., Bird, C. L., Whitby, R. J. and Day, A. E. (2013), 'First Steps Towards Semantic Descriptions of Electronic Laboratory Notebook Records', *Journal of Chemical Information and Modeling* **5**, 52.
URL: *http://www.biomedcentral.com/content/pdf/1758-2946-5-52.pdf*

Consolvo, S., Arnstein, L. and Franza, B. R. (2002), User Study Techniques in the Design and Evaluation of a Ubicomp Environment, *in* 'UbiComp 2002: Ubiquitous Computing', Springer, pp. 73–90.

Consortium, P. R. et al. (2010), 'Access vs. Importance. A Global Study Assessing the Importance of and Ease of Access to Professional and Academic information (Phase I Results)', *Retrieved from the WWW, February* **27**, 2012.

ConsortiumInfo.org (2013), Collaborative Electronic Notebook Systems Association (CENSA). [Online: Accessed 16-Jan-2018].
URL: *http://www.consortiuminfo.org/links/linksdetail.php?ID=55*

Cooke, R. and schraefel, m. c. (2004), 'Signature Flip and Clip: Virtually Flipping and Dog Earing Pages in a Digital Lab Book'.
URL: *http://eprints.soton.ac.uk/259251/*

Davis, F. D. (1989), 'Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology', *MIS Quarterly* **13**(3), 319–340.

De Waard, A. (2016), 'Research data management at Elsevier: Supporting networks of data and workflows', *Information Services & Use* **36**(1-2), 49–55.

Drăgan, L., Handschuh, S. and Decker, S. (2011), The Semantic Desktop at Work: Interlinking Notes, *in* 'Proceedings of the 7th International Conference on Semantic Systems', ACM, pp. 17–24.
URL: *http://dl.acm.org/citation.cfm?id=2063521*

Frey, J. G., De Roure, D., Mills, H., Fu, H., Peppe, S., Hughes, G., Smith, G., Payne, T. R. and others (2003), 'Context Slicing the Chemical Aether'.
URL: *http://eprints.soton.ac.uk/258790/*

Frey, J. G., Hughes, G. V., Mill, H. R., Smith, G. M., De Roure, D. and others (2004), 'Less is More: Lightweight Ontologies and User Interfaces for Smart Labs'.
URL: *http://eprints.soton.ac.uk/15882/*

Goddard, N. H., Macneil, R. and Ritchie, J. (2009), 'eCAT: Online Electronic Lab Notebook for Scientific Research', *Automated Experimentation* **1**(1), 4.
URL: *http://www.aejournal.net/content/1/1/4*

Google (2017), Colaboratory. [Online: Accessed 16-Jan-2018].
URL: *https://research.google.com/colaboratory/unregistered.html*

Greenhalgh, T. and Stones, R. (2010), 'Theorising Big IT Programmes in Healthcare: Strong Structuration Theory meets Actor-Network Theory', *Social science & medicine* **70**(9), 1285–1294.

Guerrero, S., Dujardin, G., Cabrera-Andrade, A., Paz-y Miño, C., Indacochea, A., Inglés-Ferrándiz, M., Nadimpalli, H. P., Collu, N., Dublanche, Y., Mingo, I. D. and Camargo, D. (2016), 'Analysis and Implementation of an Electronic Laboratory Notebook in a Biomedical Research Institute', *PLoS One* **11**(8).
URL: *https://search.proquest.com/docview/1808041736?accountid=13963*

Hammersley, M. and Atkinson, P. (2007), *Ethnography: Principles in Practice*, Routledge.

Hanwell, M. D., Curtis, D. E., Lonie, D. C., Vandermeersch, T., Zurek, E. and Hutchison, G. R. (2012), 'Avogadro: An Advanced Semantic Chemical Editor, Visualization, and Analysis Platform', *J. Cheminformatics* **4**(1), 17.

Hart, S. (1987), 'The Use of the Survey in Industrial Market Research', *Journal of Marketing Management* **3**(1), 25–38.

Harvey, M., Mason, N. and Rzepa, H. S. (2014), 'Digital Data Repositories in Chemistry and their Integration with Journals and Electronic Notebooks', *Journal of Chemical Information and Modeling* p. 140829132040008.
URL: *http://pubs.acs.org/doi/abs/10.1021/ci500302p*

Hawizy, L., Jessop, D. M., Adams, N. and Murray-Rust, P. (2011), 'ChemicalTagger: A Tool for Semantic Text-Mining in Chemistry', *Journal of Cheminformatics* **3**(1), 17.
URL: *http://www.biomedcentral.com/1758-2946/3/17*

Hedström, P. and Swedberg, R. (1998), *Social Mechanisms: An Analytical Approach to Social Theory*, Cambridge University Press.

Hendler, J., Shadbolt, N., Hall, W., Berners-Lee, T. and Weitzner, D. J. (2008), 'Web Science: An Interdisciplinary Approach to Understanding The Web', *Communications of the ACM* **51**(7), 60–69.

Hsieh, H.-F. and Shannon, S. E. (2005), 'Three Approaches to Qualitative Content Analysis', *Qualitative health research* **15**(9), 1277–1288.

Hughes, G., Mills, H., De Roure, D., Frey, J. G., Moreau, L., schraefel, m. c., Smith, G. and Zaluska, E. (2004b), 'The Semantic Smart Laboratory: a System for Supporting the Chemical eScientist', *Organic & Biomolecular Chemistry* **2**(22), 3284.
URL: *http://xlink.rsc.org/?DOI=b410075a*

Hughes, G., Mills, H., Smith, G., Payne, T. and Frey, J. (2004a), Breaking the Book: Translating the Chemistry Lab Book into a Pervasive Computing Lab Environment, *in* 'Proceedings of the SIGCHI Conference on Human factors in Computing Systems', ACM, pp. 25–32.

Kanare, H. M. (1985), *Writing the Laboratory Notebook*, ERIC.

Kanza, S., Willoughby, C., Gibbins, N., Whitby, R., Frey, J. G., Erjavec, J., Zupančič, K., Hren, M. and Kovač, K. (2017), 'Electronic lab notebooks: can they replace paper?', *Journal of Cheminformatics* **9**(1), 31.

Kaur, A. and Chopra, D. (2016), Comparison of Text Mining Tools, *in* '2016 5th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)', pp. 186–192.

Kihlén, M. (2005), 'Electronic Lab Notebooks–Do they work in Reality?', *Drug Discovery Today* **10**(18), 1205–1207.
URL: *http://www.sciencedirect.com/science/article/pii/S1359644605035762*

King, N. (1994), 'Qualitative Methods in Organizational Research: A Practical Guide', *The Qualitative Research Interview* p. 17.

Kipping, C. (1996), 'A Multi-Stage Approach to the Coding of Data from Open-Ended Questions', *Nurse researcher* **4**(1).

Kitzinger, J. (1995), 'Qualitative Research. Introducing Focus Groups.', *BMJ: British medical journal* **311**(7000), 299.
URL: *http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2550365/*

Kondracki, N. L., Wellman, N. S. and Amundson, D. R. (2002), 'Content Analysis: Review of Methods and their Applications in Nutrition Education', *Journal of nutrition education and behavior* **34**(4), 224–230.

Kothari, C. R. (2004), *Research methodology: Methods and techniques*, New Age International.

Latour, B. (2005), *Reassembling the Social: An Introduction to Actor-Network-Theory*, Oxford University Press.

Latour, B. (2012), *We Have Never Been Modern*, Harvard University Press.

Latour, B. and Woolgar, S. (2013), *Laboratory Life: The Construction of Scientific Facts*, Princeton University Press.

LIMSwiki (2018), ELN Vendor. [Online: Accessed 16-Jan-2018].
URL: *http://www.limswiki.org/index.php?title=ELN_vendor*

Luczak-Rösch, M. and Heese, R. (2009), Linked Data Authoring for Non-Experts, *in* 'LDOW'.

Machina, H. K. and Wild, D. J. (2013), 'Electronic Laboratory Notebooks Progress and Challenges in Implementation', *Journal of Laboratory Automation* **18**(4), 264–268.

Mack, N., Woodsong, C., MacQueen, M., K., Guest, G. and Namey, E. (2005), *Qualitative Research Methods: A Data Collectors Field Guide*, Family Health International.

Mackay, W. E., Pothier, G., Letondal, C., Bøegh, K. and Sørensen, H. E. (2002), The Missing Link: Augmenting Biology Laboratory Notebooks, *in* 'Proceedings of the 15th Annual ACM Symposium on User Interface Software and Technology', ACM, pp. 41–50.
URL: *http://dl.acm.org/citation.cfm?id=571992*

Macrina, F. L. (2005), 'Scientific Record Keeping', *Scientific Integrity* p. 269.

Manning, C. D., Raghavan, P. and Schütze, H. (2008), *Introduction to Information Retrieval*, Cambridge University Press, chapter Scoring, term weighting & the vector space model.

Mohd Zaki, Z., Dew, P. M., Lau, L. M., Rickard, A. R., Young, J. C., Farooq, T., Pilling, M. J. and Martin, C. J. (2013), 'Architecture Design of a User-Orientated Electronic Laboratory Notebook: A Case Study Within An Atmospheric Chemistry Community', *Future Generation Computer Systems* **29**(8), 2182–2196.
URL: *http://linkinghub.elsevier.com/retrieve/pii/S0167739X1300071X*

Muijs, D. (2010), *Doing Quantitative Research in Education with SPSS*, SAGE Publications.

Myers, J. D., Mendoza, E. S. and Hoopes, B. (2001), A Collaborative Electronic Laboratory Notebook, *in* 'IMSA', pp. 334–338.
URL: *http://pdf.aminer.org/000/877/371/computational_experiments_using_distributed_-tools_in_a_*
web_based_electronic.pdf

Oleksik, G., Milic-Frayling, N. and Jones, R. (2014), Study of Electronic Lab Notebook Design and Practices that Emerged in a Collaborative Scientific Environment, ACM Press, pp. 120–133.
URL: *http://dl.acm.org/citation.cfm?doid=2531602.2531709*

Palmer, J. D. and Fields, N. (1994), 'Computer Supported Cooperative Work', *Computer* **27**(5), 15–17.

Patton, M. Q. (2005), *Qualitative Research*, Wiley Online Library.

Pearce, F. T., Association, I. M. R. et al. (1966), *The Parameters of Research*, Industrial Marketing Research Association.

Punch, F, K. (2009), *Introduction to Research Methods in Education*, SAGE Publications.

Quinnell, R. and Hibbert, D. B. (2010), Introducing an Electronic Laboratory notebook to PhD Students Undertaking Chemistry Research at a Research Intensive University, *in* 'International Conference on Education, Training and Informatics: ICETI'.

Reimer, Y. J. and Douglas, S. A. (2004), 'Ethnography, Scenario-Based Observational Usability Study, and Other Reviews inform the Design of a Web-based E-notebook', *International Journal of Human-Computer Interaction* **17**(3), 403–426.

Rodden, T. (1991), 'A Survey of CSCW Systems', *Interacting with Computers* **3**(3), 319–353.

Roubert, F. and Perry, M. (2013), Putting the Lab in the Lab Book: Supporting Coordination in Large, Multi-site Research, *in* 'Proceedings of the 27th International BCS Human Computer Interaction Conference', BCS-HCI '13, British Computer Society, Swinton, UK, UK, pp. 12:1–12:10.
URL: *http://dl.acm.org/citation.cfm?id=2578048.2578066*

Rubacha, M., Rattan, A. K. and Hosselet, S. C. (2011), 'A Review of Electronic Laboratory Notebooks Available in the Market Today', *Journal of Laboratory Automation* **16**(1), 90–98.
URL: *http://jla.sagepub.com/lookup/doi/10.1016/j.jala.2009.01.002*

Rudolphi, F. and Goossen, L. J. (2012), 'Electronic Laboratory Notebook: The Academic Point of View', *Journal of Chemical Information and Modeling* **52**(2), 293–301.
URL: *http://pubs.acs.org/doi/abs/10.1021/ci2003895*

Sayre, F. D., Bakker, C., Johnston, L., Kocher, M., Lafferty, M. and Kelly, J. (2017), 'Where in Academia are ELNs? Support for Electronic Lab Notebooks at Top American Research Universities'.

schraefel, m. c., Hughes, G., Mills, H., Smith, G. and Frey, J. (2004), Making Tea: Iterative Design Through Analogy, *in* 'Proceedings of the 5th Conference on Designing Interactive Systems: Processes, Practices, Methods, and Techniques', ACM, pp. 49–58.
URL: *http://dl.acm.org/citation.cfm?id=1013124*

Shankar, K. (2004), 'Recordkeeping in the Production of Scientific Knowledge: an Ethnographic Study', *Archival Science* **4**(3-4), 367–382.

Shankar, K. (2007), 'Order from Chaos: The Poetics and Pragmatics of Scientific Recordkeeping', *Journal of the American Society for Information Science and Technology* **58**(10), 1457–1466.

Tabard, A., Mackay, W. E. and Eastmond, E. (2008), From Individual to Collaborative: The Evolution of Prism, a Hybrid Laboratory Notebook, *in* 'Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work', ACM, pp. 569–578.

Talbott, T., Peterson, M., Schwidder, J. and Myers, J. D. (2005), Adapting the Electronic Laboratory Notebook for the Semantic Era, *in* 'Proceedings of the 2005 International Symposium on Collaborative Technologies and Systems', IEEE, pp. 136–143.
URL: *http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1553305*

Tatnall, A. and Gilding, A. (2005), 'Actor-Network Theory in Information Systems Research'.

Taylor, K. R., Gledhill, R. J., Essex, J. W., Frey, J. G., Harris, S. W. and De Roure, D. C. (2006), 'Bringing Chemical Data onto the Semantic Web', *Journal of Chemical Information and Modeling* **46**(3), 939–952.
URL: *http://pubs.acs.org/doi/abs/10.1021/ci050378m*

Taylor, K. T. (2006), 'The Status of Electronic Laboratory Notebooks for Chemistry and Biology', *Current Opinion in Drug Discovery and Development* **9**(3), 348.
URL: *http://www.atriumresearch.com/library/Taylor_Electronic_laboratory_notebooks.pdf*

TechRepublic (2012), Cloud app vs. web app: Understanding the difference. [Online: Accessed 26-Jan-2018].
URL: *https://www.techrepublic.com/blog/the-enterprise-cloud/cloud-app-vs-web-app-understanding-the-differences/*

Van Vliet, H., Van Vliet, H. and Van Vliet, J. (1993), *Software Engineering: Principles and Practice*, Vol. 3, Wiley New York.

Venkatesh, V., Morris, M. G., Davis, G. B. and Davis, F. D. (2003), 'User acceptance of information technology: Toward a unified view', *MIS Quarterly* **27**(3), 425–478.
URL: *http://www.jstor.org/stable/30036540*

Voegele, C., Bouchereau, B., Robinot, N., McKay, J., Damiecki, P. and Alteyrac, L. (2013), 'A Universal Open-Source Electronic Laboratory Notebook', *Bioinformatics* **29**(13), 1710–1712.
URL: *http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/btt253*

W3C (2015), Ontologies, W3C Standard, World Wide Web Consortium. [Online: Accessed 16-Jan-2018].
URL: *https://www.w3.org/standards/semanticweb/ontology*

Walsh, E. and Cho, I. (2012), 'Using Evernote as an Electronic Lab Notebook in a Translational Science Laboratory', *Journal of Laboratory Automation* p. 2211068212471834.
URL: *http://jla.sagepub.com/content/early/2012/12/26/2211068212471834.abstract*

Weibel, J. D. (2016), 'Working toward a Paperless Undergraduate Physical Chemistry Teaching Laboratory', *Journal of Chemical Education* **93**(4), 781–784.
URL: *http://dx.doi.org/10.1021/acs.jchemed.5b00585*

Wickström, G. and Bendix, T. (2000), ''the "hawthorne effect"—what did the original hawthorne studies actually show?", *Scandinavian Journal of Work, Environment & Health* pp. 363–367.

Wolfinger, N. H. (2002), 'On Writing Fieldnotes: Collection Strategies and Background Expectancies', *Qualitative Research* **2**(1), 85–93.
URL: *http://qrj.sagepub.com/cgi/doi/10.1177/146879410200200164O*

Woolgar, S. (1997), 'The Machine at Work: Technology, Work and Organization'.

Wright, K. B. (2005), 'Researching Internet-Based Populations: Advantages and Disadvantages of Online Survey Research, Online Questionnaire Authoring Software Packages, and Web Survey Services', *Journal of Computer-Mediated Communication* **10**(3), 00–00.